CATIE CENTRO AGRONÓMICO TROPICAL DE INVESTIGACIÓN Y ENSEÑANZA

Sub-Unidad de Estadística

SAS:
Aplicaciones en el campo agropecuario y de los recursos naturales

Gustavo López
Johnny Pérez
Christoph Kleinn

Versión 1.1
Octubre, 2001

INDICE

INTRODUCCION

Este documento fue diseñado como un manual de referencia, para las personas que desean analizar datos de experimentos del campo agropecuario y forestal, utilizando como herramienta de análisis el SAS en su versión 8 para Windows. A la edición anterior a este documento titulada "Introducción al Micro SAS: Aplicaciones al análisis de experimentos agrícolas" la cual se publicó en el año de 1995, se le han agregado nuevos temas y un nuevo enfoque. Estas mejoras al documento vienen de la experiencia de los autores de trabajar asesorando a los estudiantes e investigadores del CATIE, a las sugerencias e interrogantes que los mismos usuarios han planteado en los últimos años, todo con el fin de ajustarse a las condiciones que demandan los nuevos tiempos.

Uno de los problemas que tienen los usuarios de SAS, es la carencia de documentos de referencia en español, que sean fáciles de entender por personas que no tienen conocimientos amplios en programación y estadística. Este documento trata de llenar en parte esta necesidad.

El documento se refiere al uso del SAS a través del ambiente de ventanas, por cuál se explica su uso; también cubre la programación del PASO DATA, para facilitarle al usuario la definición y el manejo de los datos y archivos que van a procesarse con SAS. También se explican los procedimientos más comunes para analizar datos de experimentos del campo agropecuario y forestal, mostrando un ejemplo, presentando el programa en SAS para el análisis e interpretando sus salidas. También se incluyen prácticas, con el objetivo de que los usuarios apliquen los conocimientos que van adquiriendo, a través de los diferentes temas que se incluyen.

Este documento se desarrolla en seis capítulos:

En el Capítulo I se explica el ambiente de trabajo del Sistema SAS, el cual permite al usuario interactuar con el software.

En el Capítulo II se explican los conceptos de archivos que puede trabajar el SAS. También se estudia la estructura de un programa SAS.

En el Capitulo III se estudian las instrucciones más importantes para una buena programación del Paso DATA y se incluyen ejemplos de programas para manipular datos y archivos.

En el Capítulo IV, se estudian algunos de los procedimientos para organizar archivos SAS, controlar las salidas de los procedimientos y realizar cálculos de estadísticas descriptivas.

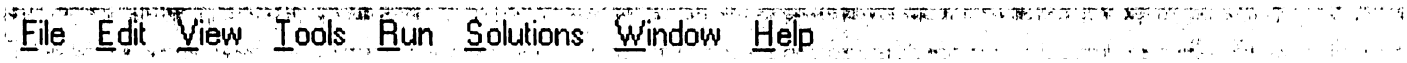
En el Capítulo V, se estudian algunos procedimientos para analizar datos de experimentos del campo agropecuario y forestal principalmente. Se dan ejemplos de los programas, sus salidas y su interpretación.

Finalmente, en el Capítulo VI se presentan otros aspectos del SAS, como es el análisis de datos interactivamente, utilizando la herramienta Analyst, la cual permite procesar datos sin necesidad de programar.

Con este documento no se pretende reemplazar los manuales de documentación del Sistema SAS. Si algún usuario quiere profundizar más en alguno de los temas desarrollados en este documento, debe consultar los manuales SAS respectivos.

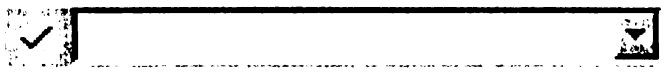
Componentes principales del Ambiente de Trabajo SAS

Barra del Menú



En esta barra se muestran distintos grupos de instrucciones relacionadas entre si, según el nombre del menú, y que se ajustan a la ventana que esta activa en el momento.

Línea de comandos



Aquí se escriben tanto comandos de ventanas como del sistema. Una vez escrito el comando se debe presionar *enter* para que este se ejecute o dar click en  .

Barra de Herramientas

Se muestran en forma de iconos (figuras) los comandos mas utilizados. Al igual que La Barra del Menú, la Barra de Herramientas se ajusta automáticamente y muestra los iconos que aplican a la ventana activa.

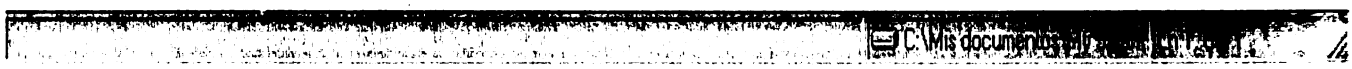
Comandos de Ventana

Sirven para minimizar, ampliar o cerrar una ventana.

Barra de Ventanas de SAS

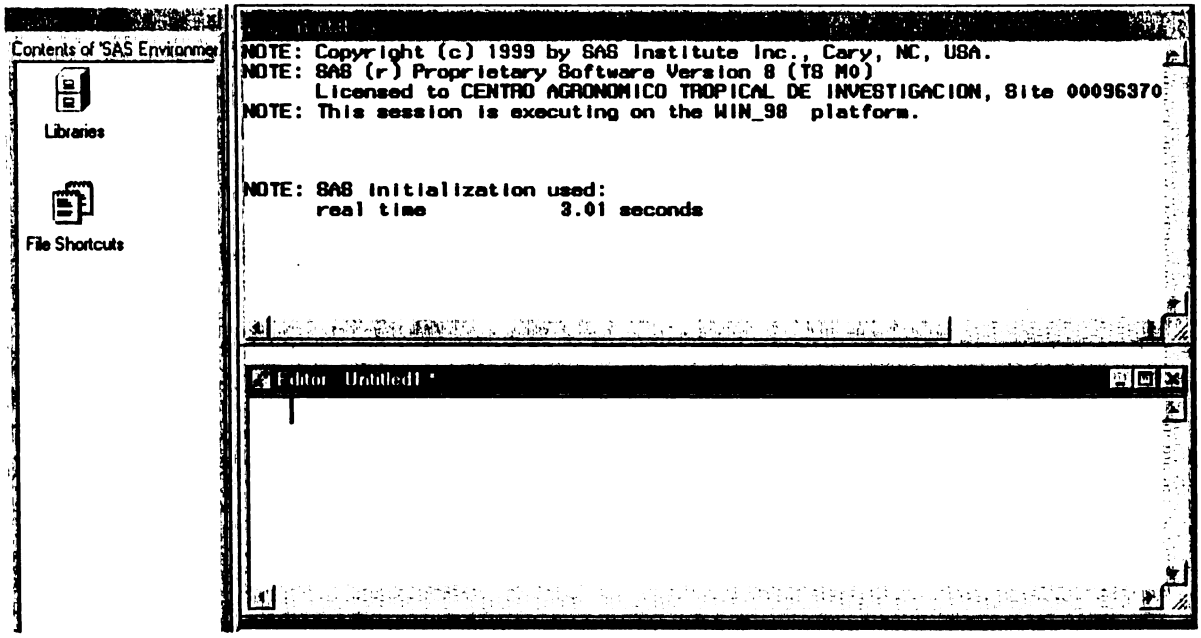
Muestra las ventanas abiertas y sirve para moverse entre ventanas. Para moverse a cualquier ventana abierta solo debe hacer un click sobre el nombre de la ventana en la Barra de Ventanas de SAS.

Barra de Mensajes y de la Carpeta Actual



En esta barra se muestran los mensajes del sistema y también muestra cual es la carpeta actual. Es una buena costumbre leer y tomar nota de los mensajes que da el sistema. Si se desea cambiar la Carpeta Actual de trabajo, solo debe darse doble click sobre el área de carpeta actual, al hacer esto se abrirá la ventana de cambio de directorio (Change Folder) en la cual deberá indicarse cual será la nueva Carpeta Actual.

Ventanas



En esta área se muestran las ventanas abiertas actualmente.

Ventanas del ambiente de trabajo SAS

Cuando se ingresa al sistema SAS, a través del ambiente de trabajo, se muestran las cinco ventanas principales del SAS: la ventana *Editor*, *Log*, *Output*, *Explorer* y *Results*.

La ventana *Editor*

Esta ventana es la que utiliza el SAS como editor. En ella se pueden escribir, editar y correr programas SAS. Tiene una serie de facilidades que le ayudaran a escribir los programas, incluyendo:

- codificación de color y revisión de la sintaxis del lengua SAS
- secciones expandibles y contraibles
- grabación de macros
- facilidades para teclas rápidas
- facilidades para deshacer y rehacer multi nivel.

En el ambiente Windows, la ventana del *Editor* aparece en lugar de la ventana *Program Editor*. Si se desea también se puede abrir esta ventana para trabajar aunque sin las facilidades de edición que brinda la ventana *Editor*.

La ventana *Log*

La ventana *Log* muestra mensajes acerca de la sesión de SAS y de cualquier programa SAS que se haya ejecutado. La información que el sistema despliega en esta ventana incluye: mensajes de error, características sobre los archivos que trabaja el programa (nombre, variables, observaciones) y tiempos de duración en la ejecución de las diferentes partes del programa.

La ventana *Output*

En esta ventana se despliegan los resultados (salidas) que producen los diferentes procedimientos utilizados en el programa que se ejecuta. Si el programa consiste solo de pasos DATA en esta ventana no se despliega nada. Por defectos esta ventana se encuentra detrás de las otras ventanas. Cuando se crea una salida, la ventana *Output* automáticamente de presente al frente de la pantalla.

La ventana *Explorer*

En la ventana *Explorer*, se pueden ver y manejar los archivos SAS, y crear accesos rápidos (shortcuts) para archivos que no sean SAS. En esta ventana también se pueden crear nuevas librerías y archivos SAS, abrir cualquier archivo SAS, y realizar mas tareas para el manejo de archivos tales como mover, copiar y borrar archivos.
Se puede hacer que la ventana del *Explorer* muestre su contenido en forma de árbol.

La ventana *Results*

La ventana *Results* ayuda a moverse y a manejar las salidas de los programas que se hayan ejecutado. En esta se pueden ver, gravar e imprimir individualmente partes de las salidas.

Por defecto, la ventana *Results* se muestra detrás de la ventana *Explorer* y esta se encuentra vacía hasta que la ejecución de un programa SAS genere una salida. Entonces cuando esto sucede, esta se mueve al frente de la pantalla conjuntamente con la ventana *Output*.

Otras ventanas

Además de las cinco ventanas principales existen muchas otras ventanas, las cuales completan las distintas aplicaciones, componentes y facilidades del sistema SAS. Entre estas podemos citar las siguientes ventanas:

KEYS: Para ver, incluir o modificar las teclas rápidas de funciones programadas. Aparte de las teclas rápidas que por defecto trae el sistema, se pueden configurar a gusto del usuario. Para activar esta ventana seleccione en la barra del menú **Tools ➡ Options ➡ Keys**

HELP: Para obtener información general sobre cualquier tema del sistema SAS. Para activar esta ventana seleccione el menú **Help**. También dando click en el icono  se obtendrá la ayuda de la ventana activa o del tema relacionado directamente.

ANALYST: Esta es una herramienta de análisis diseñada para proveer un fácil acceso para un análisis estadístico básico. Esta aplicación esta orientada a estudiantes y usuarios nuevos de SAS, aunque también facilitara el análisis para investigadores y usuarios expertos del SAS. Para entrar al Analyst seleccione en la barra del menú **Solutions ➡ Analysis ➡ Analyst**.

ASSIST: Es una interfase del sistema SAS para dar asistencia para la realización de tareas. Con esta se puede tener acceso a todo el poder del SAS, sin tener que conocer los comandos de programación del SAS. Para ingresar al Assist seleccione en la barra del menú **Solutions ➡ ASSIST**.

Activación de una ventana

La ventana activa es aquella donde se encuentra el cursor actualmente. En la ventana activa se pueden emitir comandos del sistema; en el *Editor* y en el *Program Editor*, además se puede entrar texto y ejecutar programas.

En el ambiente de ventanas, solo una está activa a la vez. Al entrar al sistema SAS, la ventana activa es la del *Editor*.

Se puede activar una ventana de las siguientes formas:

- Dar un click en cualquier área de la ventana que se desea activar.
- Entrar el nombre de la ventana que se quiere activar en la línea de comandos de la ventana activa.

- Usar las teclas de funciones que corresponden a cada ventana. Las teclas son las siguientes:

F1	Activa la ventana Help
F5	Activa la ventana PROGRAM EDITOR
F6	Activa la ventana LOG
F7	Activa la ventana OUTPUT

- Dar un click sobre el nombre de la ventana en la Barra de Ventanas de SAS.

- En el menú Windows seleccionar de la lista de ventanas abiertas. la ventana que se desea activar.

Almacenamiento de ventanas

El usuario puede almacenar el contenido de las ventanas primarias (Editor o PROGRAM EDITOR, LOG, OUTPUT) en un archivo. Lo mas común es gravar en formato de texto, aunque también se puede gravar en formato HTML o RTF.

Para almacenar el contenido de cualquiera de estas ventanas, se debe primero activar la ventana y si esta contiene alguna información, entonces abrir y completar los campos de la ventana de gravar (Save As). Para esto, de un click al icono de gravar  (Save), o valla al menú **File ➡ Save** (o **Save As**) y especifique la carpeta, el nombre y el formato (tipo) del archivo que desea gravar.

El nombre del archivo debe tener las características de archivos Windows, con un nombre (máximo 256 caracteres) y una extensión (máximo 3 caracteres).

Para una mejor documentación y orden con los archivos creados por las ventanas SAS se recomiendan las siguientes extensiones:

	Para almacenar el contenido de la ventana OUTPUT
	Para almacenar el contenido de la ventana LOG.
	Para almacenar el contenido de la ventana Editor o PROGRAM EDITOR, si este es un programa SAS.
	Para almacenar el contenido de la ventana editor o PROGRAM EDITOR, si este es un conjunto de datos.

Cargando archivos en el Editor

Tanto en la ventan Editor como en la ventana Program Editor, se pueden llamar archivos externos en formato de texto (ASCII). Generalmente estos serán programas SAS gravados con anterioridad. Para cargar estos archivos a estas ventanas, primero active alguna de estas ventanas, después vaya al menú **File ➡ Open** y complete los campos de la ventana de Abrir (Open). Otra opción es con el icono  (Open) de la barra de herramientas.

Impresión de ventanas

Para imprimir el contenido de una ventana, primero se activa, luego se ingresa al menú **File ➡ Print**, o desde la barra de herramientas se selecciona el icono  (Print).

Ejecución de un programa SAS

Para ejecutar un programa a través del ambiente de trabajo SAS, este debe estar cargado en la ventana del Editor o Program Editor. La ejecución se puede realizar de varias formas:

- Desde el menú **Run** seleccionar el comando **Submit**.
- Digitando el comando SUBMIT en la línea de comandos.
- Oprimiendo la tecla F8.
- Dar click al icono  en la barra de herramientas.

Cuando el programa se inicia, y si la ventana del Editor no ocupa toda el área de las ventanas, se observará en forma muy rápida como se ejecutan los comandos en la ventana LOG y si el programa termina con éxito, automáticamente se mostrara la ventana OUTPUT sobre todas las otras ventanas y también se mostrara la ventana Results, en la cual podrá desplazarse directamente a la parte de la salida que le interese, seleccionándola de la lista. Si el programa no se ejecuta con éxito o este no produce ninguna salida, la ventana OUPUT no se mostrará y aparecerá la ventana Editor como la ventana activa al finalizar la ejecución del programa.

El SAS despliega en la ventana LOG todos los mensajes sobre la ejecución del programa. Algunos de estos mensajes son: de errores, contenido de los archivos, tiempo de ejecución, uso de memoria, etc.
En la ventana OUTPUT el sistema despliega todas las salidas de los diferentes procedimientos que incluye el programa. Si el programa tiene solo pasos DATA (sin ninguna instrucción PROC), la ventana OUTPUT queda vacía.

Lectura de mensajes y corrección de programas

Después que un programa termina de ejecutarse, se debe activar la ventana LOG para ver los mensajes sobre la ejecución de este. Si hay errores en el programa, se deben seguir los siguientes pasos:

- Tomar nota de los errores en la ventana LOG.
- Limpiar las ventanas LOG y OUTPUT. Para hacerlo digite el comando CLEAR en la línea de comandos de cada ventana.
- Activar la ventana Editor. Si esta trabajando con la ventana PROGRAM EDITOR, recuperar el programa digitando el comando RECALL desde la línea de comandos o presionando la tecla F4.
- Hacer las correcciones necesarias en el programa.
- Ejecutar el programa corregido.

Estos pasos deben realizarse las veces que sea necesario hasta que el programa funcione sin errores.

PRACTICA

Digite el siguiente programa SAS en la Ventana EDITOR. Ejecute el programa y grabe las ventas OUTPUT, LOG, EDITOR.

```
DATA Ejemplo;
INPUT Rep Trat Nplanpa Rpa;
LABEL Rep='Repetición'
      Trat='Tratamiento'
   Nplapa='Número plantas parcela'
      Rpa='Rendimiento parcela';
CARDS;
1 1 56   8.5
2 1 65  12.5
1 2 48   6.3
2 2 36   5.2
1 3 74  14.8
2 3 85  16.3
;
PROC PRINT LABEL SPLIT=' ';
RUN;
```

Si tiene errores, revisar la Ventana LOG, corregir los errores y volver a ejecutar el programa

CAPITULO II: Archivos y Estructura de un Programa SAS

Archivos de Trabajo SAS

Para procesar datos con SAS, es necesario que estos estén organizados y grabados como archivos de trabajo SAS. Un archivo de trabajo SAS está representado por una matriz bidimensional de filas y columnas, donde las filas representan las observaciones (unidades de investigación) y las columnas las variables (características medidas). La siguiente tabla representa la estructura de un archivo de trabajo SAS:

Nombre	Sexo	Edad	Estatura	Peso
Juan	M	33	173	170
María	F	25	169	125
Luis	M	19	180	205
Carlos	M	27	163	140
Ana	F	23	176	136
José	M	31	183	179

En la tabla anterior, cada una de las columnas: nombre, sexo, edad, estatura y peso representan las variables del archivo de trabajo SAS; y cada fila (una persona) representan las observaciones del archivo de trabajo SAS. El archivo tiene 5 variables y 6 observaciones.

A través de un programa SAS se puede tener acceso a los siguientes tipos de archivos para generar archivos de trabajo SAS:

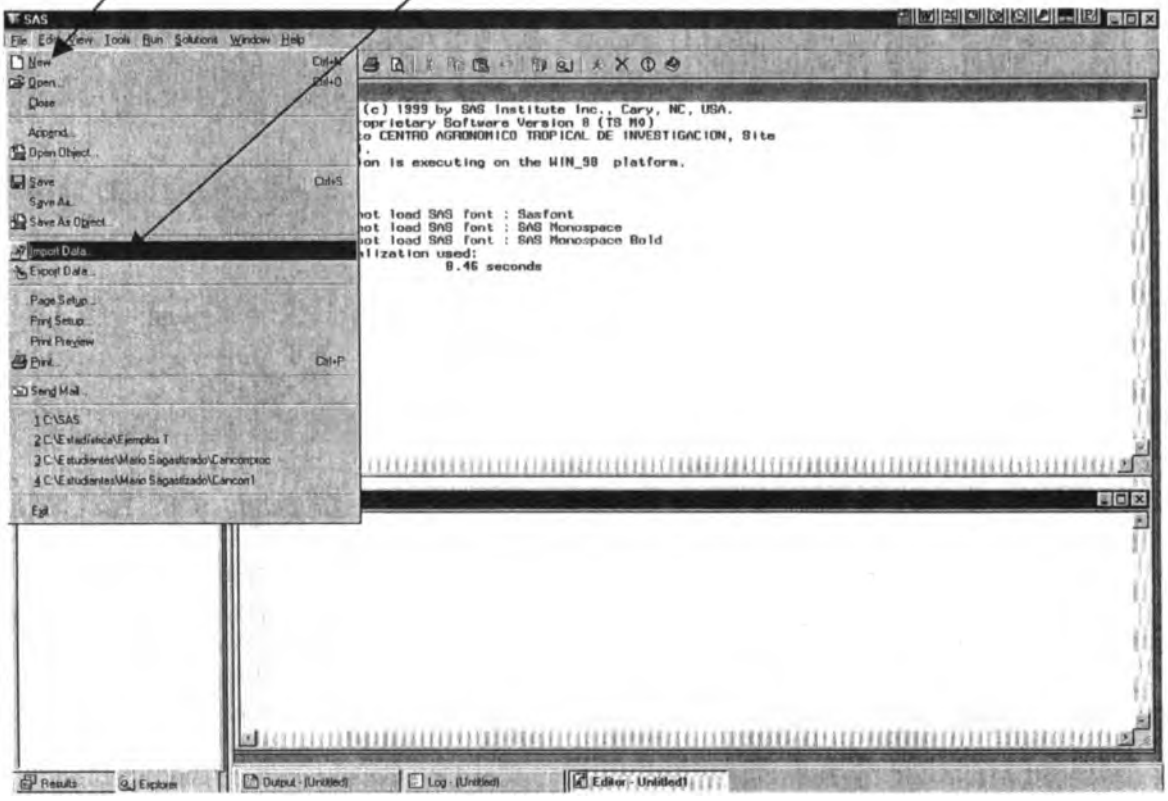
- Archivos de texto (ASCII), los cuales pueden ser generadores por la ventana EDITOR del SAS, Excel, Word, otros editores, etc.
- Archivos DBF (bases de datos), los cuales pueden ser generados por DBase, FoxPRo, ACCES, Visual Basic, etc.
- Archivos de trabajo SAS, los cuales son generados por el paso DATA de SAS

También existe el PROC IMPORT, el cuál importa directamente datos en otros formatos (Excel, Lotus, Dbase) a archivos de trabajo SAS. Esta importación se puede también realizar con la opción Import Data del menú File del SAS.

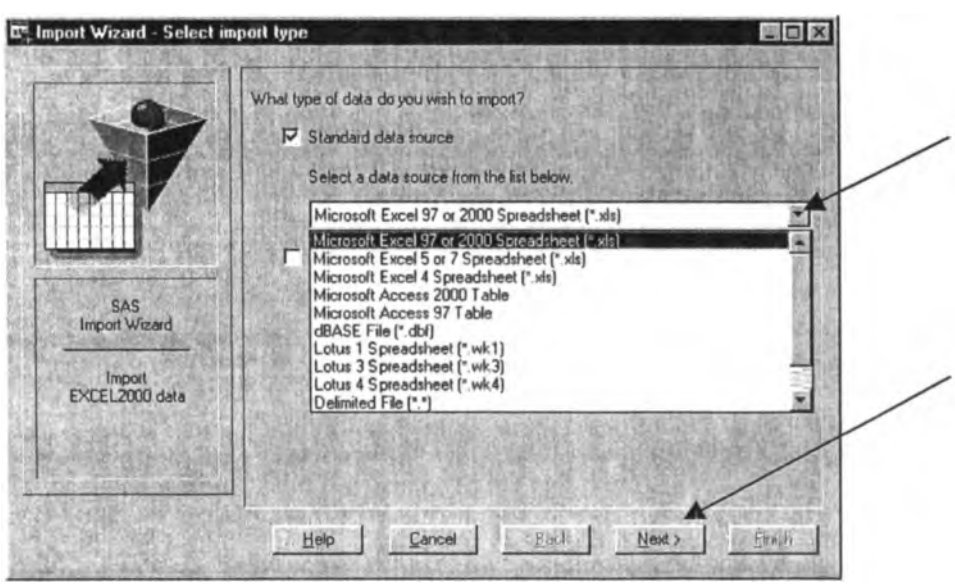
A continuación se presenta la metodología para la importación de datos:

- Entrar al programa SAS.

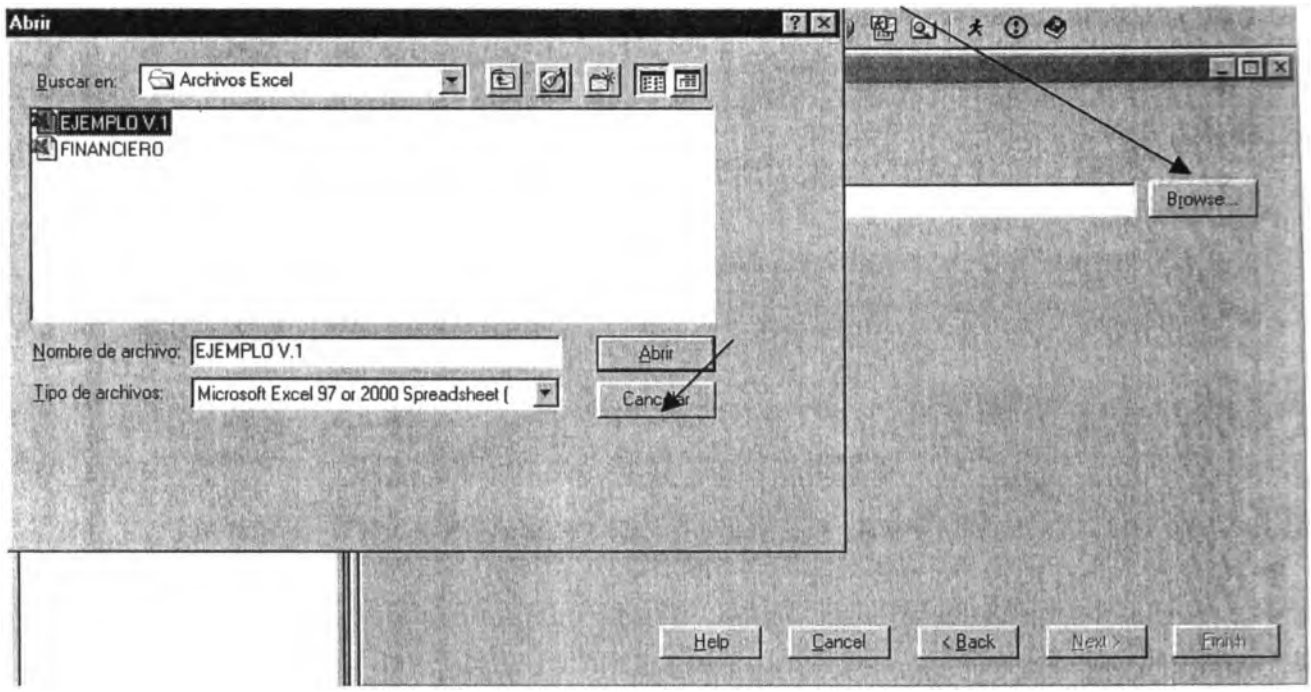
- Ir a **File** y elegir la opción **Import Data**.



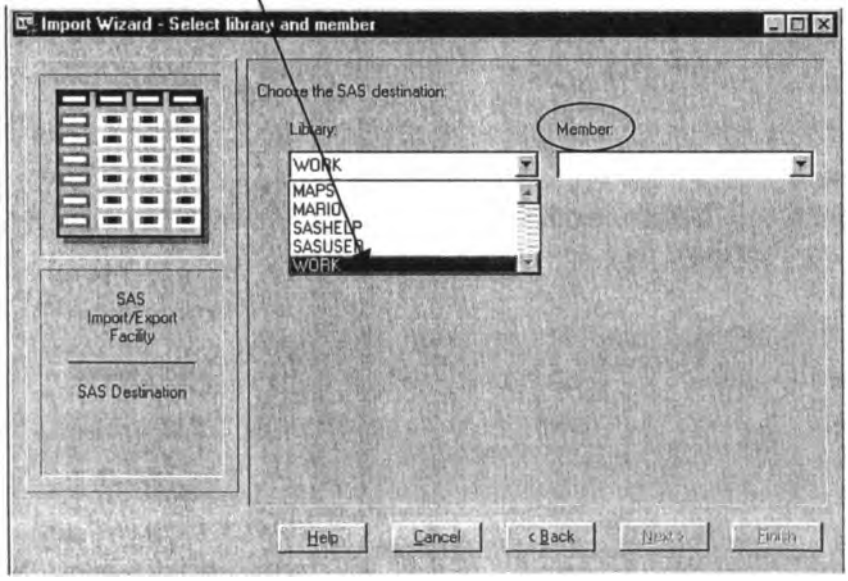
- **Inmediatamente se va desplegar una ventana en la cuál hay que seleccionar la Versión del Microsoft Excel, Microsoft Access, Archivos dbf o Lotus del cuál se van a importar los datos. Una vez elegido el Programa y la Versión oprimir Next.**



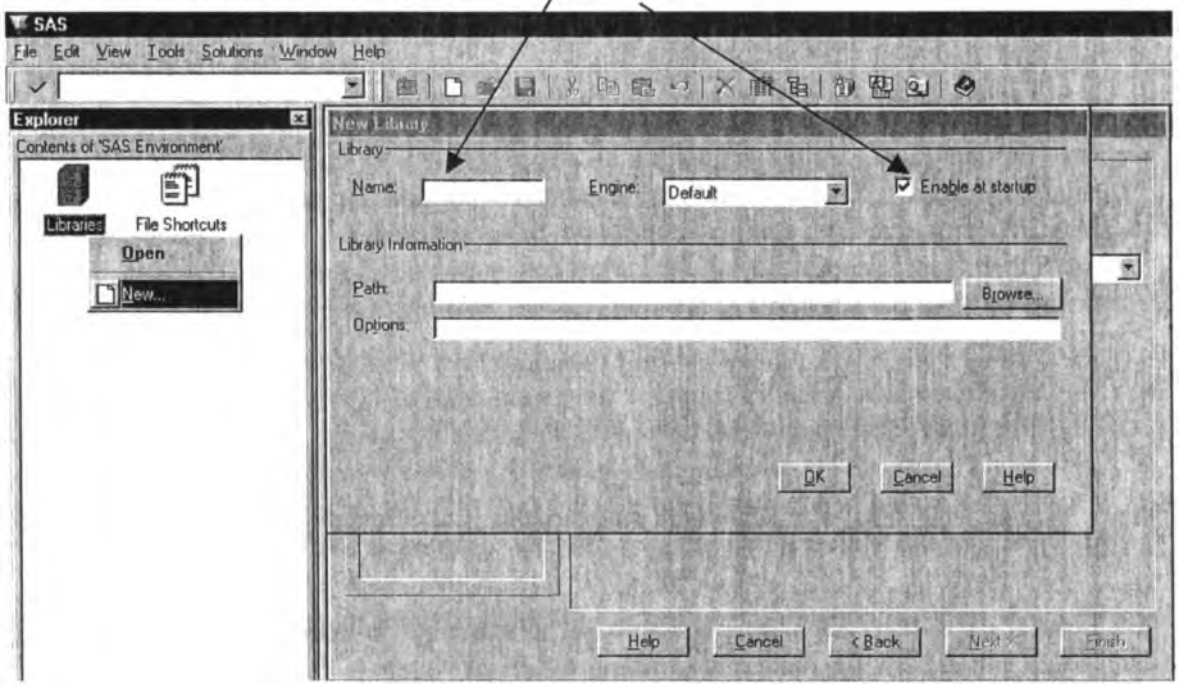
Ubicar el archivo que va a importar. Para ello seleccionar Browse y Abrir.



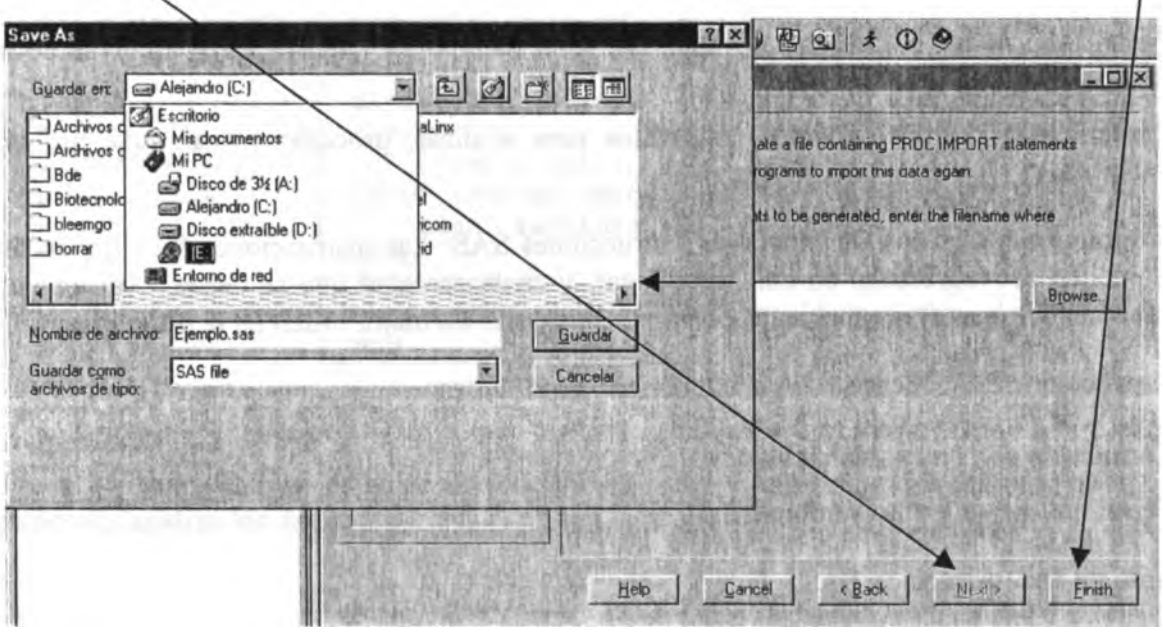
Seleccionar la librería en donde va a guardar el archivo y darle un nombre al mismo. El SAS posee una librería llamada WORK que es de tipo temporal. También se puede guardar el archivo en una librería permanente . Esta librería tiene que ser generada con anterioridad.



Para crear una librería se posiciona en el icono LIBRARY y oprime CLICK derecho seleccionando la opción NEW, que va desplegar una ventana en la que generalmente se le da nombre a la librería, y además se marca la opción ENABLE AT STARTUP.



Posteriormente después de haber elegido la librería en donde se va a guardad el archivo, se oprime NEXT , pasando a una última ventana que si se quiere puede omitirse, ya que es para guardar el archivo (que sería ya de tipo SAS) en una carpeta determinada. Por último seleccionar FINISH para así obtener la importación de los datos.



Tipos de Archivos de Trabajo SAS

Existen dos tipos de archivos de trabajo SAS:

- **Temporales:** son aquellos creados en un paso DATA, con los cuales se pueden utilizar procedimientos SAS para analizar los datos sólo durante el transcurso de una sesión de SAS. Al finalizar la sesión, el archivo SAS se borrará. Si se desean realizar análisis posteriores, será necesario volver a entrar a una sesión SAS y escribir nuevamente el programa SAS para generar el archivo.

- **Permanentes:** Son aquellos creados en un paso DATA, con los cuales se pueden utilizar procedimientos SAS para analizar los datos durante la sesión SAS. Al finalizar la sesión, el archivo de trabajo SAS permanecerá grabado y posteriormente puede volver a ser utilizado sin necesidad de ejecutar de nuevo el paso DATA. Un archivo de trabajo SAS permanente puede ser procesado directamente por los procedimientos SAS.

Un archivo de trabajo SAS será temporal o permanente de acuerdo a como lo especifique el usuario en el paso DATA (esto se estudiará posteriormente)

Estructura de un Programa SAS

Todos los programas SAS son una secuencia de pasos SAS. Existen solamente dos clases de pasos SAS:

- **Pasos DATA**, los cuales graban, leen, corrigen, manipulan y transforman archivos de trabajo SAS
- **Pasos PROC** (Procedimientos) utilizados para analizar, procesar y manipular archivos de trabajo SAS

Un programa SAS es una secuencia de instrucciones SAS. Las instrucciones SAS tienen formato libre, pueden ser ingresadas en cualquier lugar, y en la cantidad que se desee. Sin embargo, es recomendable que se digite una instrucción por línea para un mejor orden en la programación.

Las instrucciones SAS tienen dos características importantes:

- Comienzan con una palabra clave y
- Terminan con un punto y coma (;)

Ejemplo de una instrucción SAS:

```
INPUT  Nombre $ Sexo $ Edad Estatura Peso;
```

Un programa SAS puede estar compuesto de:

- **Sólo pasos DATA.** Por ejemplo, cuando se quiere únicamente generar un archivo de trabajo SAS permanente, para posteriormente analizarlo. En este caso los datos fuente, casi siempre van a estar en archivos de texto o archivos DBF (Ver Figura 1)

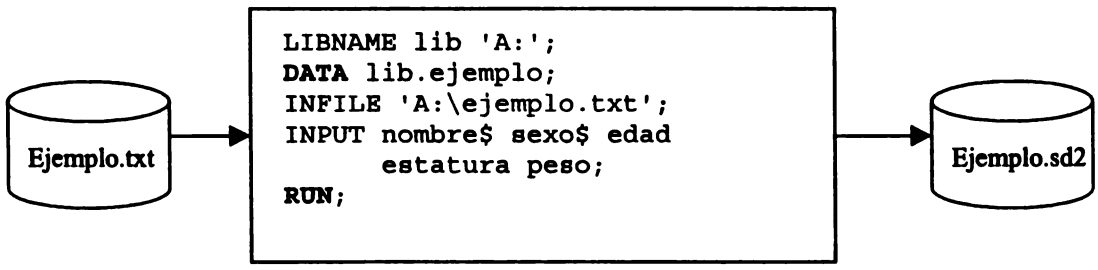


Figura 1. Programa que lee un archivo de texto (ejemplo.txt) y crea un archivo de trabajo SAS permanente (ejemplo.sd2)

- **Combinaciones de pasos DATA y pasos PROC.** Por ejemplo, cuando se leen archivos en formato de texto, se crean archivos de trabajo SAS y se analizan los datos. (Ver Figura 2)

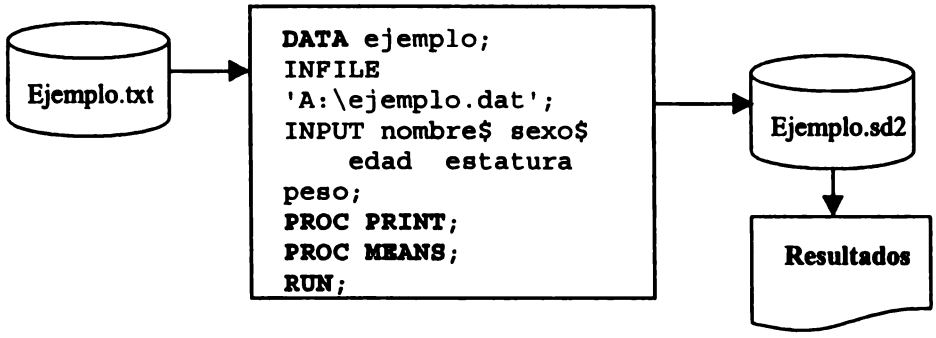


Figura 2. Programa que lee un archivo de texto (ejemplo.txt), crea un archivo de trabajo SAS temporal y analiza los datos.

- **Sólo pasos PROC**. Cuando se quiere analizar datos que ya están en archivos de trabajo SAS permanentes. (Ver Figura 3)

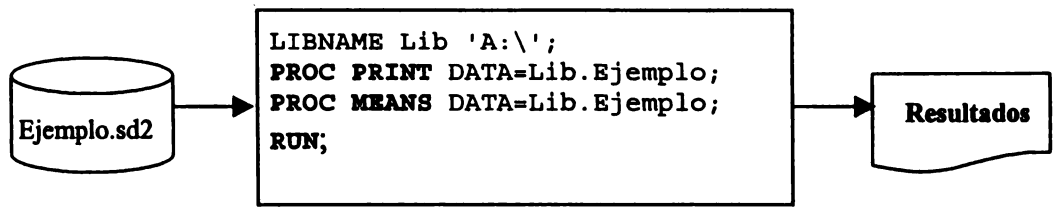


Figura 3. Programa que lee un archivo de trabajo SAS permanente (ejemplo.sd2) y procesa los datos.

Resumen

Los datos que se van a analizar con SAS deben estar grabados como archivos de trabajo SAS (formato SAS)

Un archivo de trabajo SAS es una matriz bidimensional, donde las filas representan las observaciones y las columnas las variables.

Los datos fuente o crudos que se pueden cargar a un archivo de trabajo SAS, pueden estar en formato de texto, formato dbf (base de datos) o archivos de trabajo SAS. Un archivo de trabajo SAS puede definirse como temporal o permanente, siendo la única diferencia, que el archivo permanente se conservará en el directorio especificado al terminar la sesión SAS.

Un programa SAS puede ser una combinación de pasos DATA y pasos PROC. Un paso DATA, es un conjunto de instrucciones que se utilizan para definir los datos y un paso PROC, es un conjunto de instrucciones que se utilizan para analizar los datos.

Como cualquier lenguaje, la programación SAS tiene una sintaxis que respetar. Lo más relevante de ésta, es que cada instrucción inicia con una palabra clave (palabras reservadas por el SAS) y terminan con un punto y coma (;)

CAPITULO III: Programación del paso DATA

El Paso DATA

El paso DATA es una serie ordenada de instrucciones para manejar conjuntos de datos, que comienza con la instrucción DATA. Las principales funciones de este paso son:

- Leer archivos externos y crear archivos de trabajo SAS
- Actualizar archivos SAS ya existentes
- Unir archivos de trabajo SAS: concatenar horizontal o verticalmente
- Seleccionar o eliminar registros y/o variables
- Transformar variables
- Generar nuevas variables
- Desarrollar aplicaciones propias (programación pura)

El conjunto de datos que se procesa en un paso DATA puede corresponder a alguna de las siguientes situaciones:

- **Datos grabados en archivos en formato de texto**; cuando se presenta esta situación, el paso DATA debe incluir al menos las siguientes instrucciones:

```
DATA nombre;
INFILE 'A:\nombre.dat';
INPUT lista de variables;
```

- **Datos incluidos en el programa.** Cuando los datos por procesar se van a incluir en el programa, se deben digitar al menos las siguientes instrucciones:

```
DATA nombre;
INPUT lista de variables;
CARDS;
líneas de datos
; (fin de datos)
```

- **Datos grabados en formato DBF** (bases de datos). Cuando los datos por procesar están grabados en archivos DBF, se debe incluir al menos las siguientes instrucciones:

```
FILENAME nombre 'A:\ejemplo.dbf';
PROC DBF DB4=nombre OUT=ejemplo;
```

- **Datos grabados en formato Excel u otros tipos de formatos.** Cuando los datos están en archivos Excel u otro tipos de formatos, como Lotus, Acces o DBF's se puede utilizar el PROC IMPORT, el cual importa directamente el archivo a formato SAS. Lo más común es que los usuarios de SAS graben sus datos con Excel, esto debido a la facilidad de este software para entrar datos. A continuación un ejemplo del PROC IMPORT, para un archivo Excel.

```
PROC IMPORT
   OUT=Ejemplo
   DATAFILE= "C:\ejemplo.xls"
   DBMS=EXCEL2000 REPLACE;
   GETNAMES=yes;
RUN;
```

  En el programa anterior, se esta generando un archivo de trabajo SAS temporal cuyo nombre es Ejemplo (OUT)

  Los datos fuente, están en el archivo ejemplo.xls en el directorio C. (DATAFILE)

  El formato de los datos es Excel 2000 (DBMS)

  Se utilizó en Excel la primera fila para indicar los nombres que van a tener las variables en el archivo de trabajo SAS. (GETNAMES=yes)

  La importación de datos también se puede realizar con la opción **Import Data**, del menú **File** del SAS. Esta opción le presentará una serie de ventanas para la importación de los datos. Si usa este modo, no es necesario utilizar el PROC IMPORT.

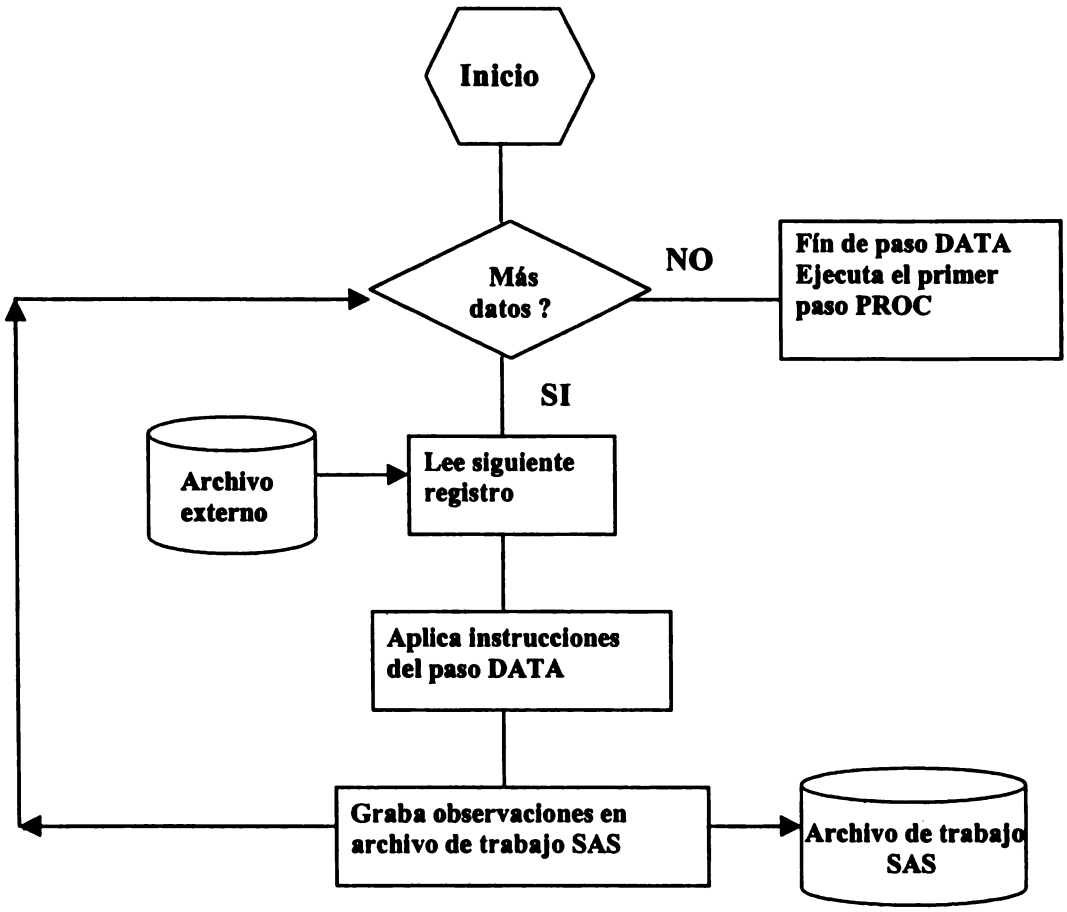
- **Datos existentes en archivos de trabajo SAS.** Cuando los datos por procesar están grabados en archivos de trabajo SAS, se deben incluir al menos las siguientes instrucciones:

```
DATA nombre;
 SET [MERGE] nombre;
```

La forma en que procede el paso DATA para generar el archivo SAS se ilustra en la figura 4. Este paso incluye al menos las siguientes instrucciones.

- **DATA** indica el comienzo de un paso DATA y a continuación se escribe el nombre del archivo de trabajo SAS.
- **INPUT** se indican las variables que contiene el conjunto de datos que se va a trabajar.
- **INFILE** se indica el directorio y el nombre del archivo que contienen los datos que se grabarán en el archivo de trabajo SAS.

Figura 4. Funcionamiento del paso DATA para la creación del archivo de trabajo SAS .



La Instrucción DATA

En la instrucción DATA, se da el nombre del archivo SAS que se quiere generar. El nombre debe tener como máximo 32 dígitos, los cuales pueden ser letras o números, pero el primero debe ser siempre una letra y no se permiten espacios en blanco. También le indica a SAS que se inicia un paso DATA. Ejemplos:

```
DATA Ejemplo; * genera un archivo de trabajo SAS  temporal;
DATA Lib.Ejemplo; *genera un archivo de trabajo SAS permanente;
```

En un programa SAS, en el nombre de un archivo de trabajo temporal se indica un sólo nivel (una sola palabra). En el nombre de un archivo de trabajo permanente se indican dos niveles (dos palabras separadas por un punto).

La Instrucción LIBNAME

El sistema SAS utiliza la instrucción LIBNAME para definir el directorio donde van a quedar almacenados los archivos SAS permanentes; es necesario incluir la instrucción LIBNAME junto al paso DATA (antes de la instrucción DATA) para crear o accesar el archivo permanente. En la instrucción LIBNAME se da el nombre (cualquier nombre, que sirve como sobrenombre (alias) temporal, de ocho caracteres o menos) del directorio donde se grabará el archivo. Después que termina la sesión SAS, el archivo permanente continúa almacenado en el directorio indicado en la instrucción LIBNAME. **Ejemplo:**

```
LIBNAME Lib 'C:\DATOS';
DATA Lib.Ejemplo;
```

En la instrucción LIBNAME anterior, se le indica a SAS que el archivo permanente se grabe en el directorio DATOS del disco duro C. El primer nombre del archivo en la instrucción DATA debe ser igual al nombre especificado en la instrucción LIBNAME, que en este caso es "Lib". Cuando se ejecute el programa en la sesión SAS, el nombre del archivo para SAS será Lib.Ejemplo. Al finalizar la sesión, el nombre del archivo para el Sistema Operativo será Ejemplo.SAS7BDAT. La extensión SAS7BDAT la define automáticamente el Sistema Operativo para todos los archivos de trabajo SAS.

Lo que el usuario debe tener presente, es donde físicamente reside el archivo de trabajo y el nombre. A continuación se muestran las instrucciones que procesan un archivo de trabajo SAS permanente.

```
LIBNAME Lib 'C:\DATOS';
PROC PRINT DATA=Lib.Ejemplo;
PROC MEANS DATA=Lib.Ejemplo;
RUN;
```

Este pequeño programa imprime los datos y calcula estadísticas descriptivas a un archivo de trabajo SAS permanente que se llama Ejemplo.SAS7BDAT y que se encuentra grabado en el directorio .

La Instrucción INFILE

La instrucción INFILE se utiliza cuando el programa va a procesar datos que se encuentran en archivos externos (formato de texto). En la instrucción se indica el directorio y el nombre del archivo externo que se va a leer. **Ejemplo:**

```
INFILE  'C:\CURSO\Ejemplo.txt';
```

La instrucción CARDS

Si el conjunto de datos que se va a procesar se incluye dentro del programa, la instrucción CARDS indica que los datos se incluyen a continuación. Cuando se utiliza la instrucción CARDS, ésta debe ser la última instrucción del paso DATA y debe aparecer inmeditamente antes de los datos. **Ejemplo:**

```
DATA nombre;
INPUT lista de variables;
* Otras instrucciones del paso DATA;
CARDS;
líneas de datos
; * (indica el fín de los datos);
```

La instrucción INPUT

La instrucción INPUT se utiliza para:

- Asignar nombre a las variables del archivo de trabajo
- Indicar el tipo de la variable: númerica, de caracter, de fecha, de tiempo
- Indicar la forma (formato) en que están organizados los datos
- Leer las líneas de datos que se grabarán en el archivo de trabajo

La instrucción INPUT tiene tres formas, de acuerdo al modo como estén organizados los datos fuente:

- columna
- lista
- formato

En una instrucción INPUT se pueden mezclar especificaciones correspondientes a estas formas. Además, para flexibilizar aún más la entrada de datos, se pueden utilizar "controles del apuntador". Un apuntador es un indicador que se posiciona en una columna de un registro o se mueve de un registro a otro.

La instrucción INPUT permite controlar el desplazamiento del apuntador, como se verá más adelante.

Entrada de Datos en Formato de Columna

Se especifica dónde encontrar los valores en el registro de entrada por medio de las posiciones de las columnas.

Ejemplo: se tiene el siguiente conjunto de datos en formato de columna:

NOMBRE									Sexo	Edad		Altura			Peso		
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
J	U	A	N		J	O	S	E		M	3	3	1	7	3	1	7	0
A	L	E	J	A	N	D	R	A		F	2	5	1	6	9	1	2	5
F	E	R	N	A	N	D	O			M	1	9	1	8	0	2	0	5
G	U	I	L	L	E	R	M	O		M	2	7	1	6	3	1	4	0
P	A	T	R	I	C	I	A			F	2	3	1	7	6	1	3	6
R	I	G	O	B	E	R	T	O		M	3	1	1	8	3	1	7	9

La tabla anterior contiene las variables: nombre, sexo, edad, altura y peso. Los datos se encuentran en formato de columna; para cada variable se reservan columnas especificadas que se mantienen constantes para todas las observaciones. La instrucción INPUT para este conjunto de datos es la siguiente:

```
INPUT  Nombre $ 1-9 Sexo $ 11 Edad 12-13 Altura 14-16 Peso 17-19;
```

Note que en esta instrucción, las variables Nombre y Sexo van seguidas del signo ($), para indicar que la variable es alfabética. Las demás variables: Edad, Altura y Peso son numéricas. Como se muestra en la instrucción INPUT en formato de columna, se da el nombre a la variable (máximo 32 dígitos. Deben empezar siempre con una letra y no se permiten espacios en blanco) y se indica el tipo de la variable (si es numérica no se indica nada), además de la columna inicial y final donde se encuentra el dato.

Cuando se utiliza este formato, las posiciones de las variables deben ser fijas; es decir, si en el INPUT se indica que la variable Nombre está en las columnas 1 a 9, los datos para esa variable deben estar siempre entre esas posiciones. Este tipo de formato permite leer todo o parte de los datos. En este ejemplo se lee todo el registro. Sin embargo, es posible leer sólo algunas variables; por ejemplo, se se desea leer sólo el nombre y la edad, la instrucción sería:

```
INPUT Nombre $ 1-9 Edad 12-13;
```

Entrada de Datos en Formato Libre

Las variables se listan en el orden en el que aparecen en el registro de entrada. Se deben tener en cuenta los siguientes aspectos en este formato de entrada de datos.

- El orden en el que aparecen las variables en la instrucción INPUT debe coincidir exactamente con el orden de las variables en el archivo por leer.
- Los valores de las variables deben estar separados al menos por un espacio en blanco.
- En los valores para las variables alfabéticas no se permiten espacios intermedios. La longitud máxima es de ocho dígitos.
- Si se dejan espacios en blanco, cuando falta el valor de una variable en una observación, los nombres de las variables y sus valores se desfazan. Los valores faltantes en este tipo de formato se deben indicar con un punto (.).
- Las posiciones de las variables no tienen que ser fijas

Ejemplo: se tiene el siguiente conjunto de datos en formato de Lista:

NOMBRE									Sexo	Edad			Altura				Peso			
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
J	U	A	N	J	O	S	E			M		3	3		1	7	3		1	7	0
A	L	E	J	A	N	D	R	A		F		2	5		1	6	9		1	2	5
F	E	R	N	A	N	D	O			M		1	9		1	8	0		2	0	5
G	U	I	L	L	E	R	M	O		M			.		1	6	3		1	4	0
P	A	T	R	I	C	I	A			F		2	3		1	7	6		1	3	6
R	I	G	O	B	E	R	T	O		M		3	1		1	8	3		1	7	9

La instrucción INPUT para este conjunto de datos es de la siguiente forma:

```
INPUT Nombre $ Sexo $ Edad Altura Peso;
```

En el cuadro anterior se observa que:

- Existe al menos un espacio en blanco para separar los valores de las variables.
- Los valores de las variables guardan un orden entre una columna inicial y una columna final. Esto no es necesario; sin embargo, para una revisión posterior de los datos, esta forma puede facilitar la tarea.

- En la línea que corresponde a GUILLERMO, no se tiene el dato para la variable Edad. Se digita un punto (.) para indicar que es un "valor faltante". Cuando se ejecuta algún procedimiento, este no considera el "valor faltante" al momento de realizar los cálculos.

- La variable nombre, en algunos de sus datos, tiene más de ocho dígitos. En estos casos solamente se registran hasta ocho caracteres, por lo cual algunos nombres quedan truncados (Ver instrucción LENGTH para manejar este problema).

Entrada de Datos con Formato

Se especifica la longitud de la variable y si tiene decimales o no. Los formatos que se tienen son los siguientes:

w. Longitud (ancho) del campo numérico, sin decimales

w.d Numérico con decimales

$w. Longitud del campo de caracteres

Donde w y d simbolizan números.

Ejemplo: se tiene el siguiente conjunto de datos en forma de formato:

NOMBRE									SEXO	EDAD		ALTURA				PESO		
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
J	U	A	N		J	O	S	E	M	3	3	1	.	7	3	1	7	0
A	L	E	J	A	N	D	R	A	F	2	5	1	.	6	9	1	2	5
F	E	R	N	A	N	D	O		M	1	9	1	.	8	0	2	0	5
G	U	I	L	L	E	R	M	O	M		.	1	.	6	3	1	4	0
P	A	T	R	I	C	I	A		F	2	3	1	.	7	6	1	3	6
R	I	G	O	B	E	R	T	O	M	3	1	1	.	8	3	1	7	9

La siguiente instrucción INPUT lee este conjunto de datos con formato:

```
INPUT Nombre $ 9. Sexo $1. Edad 2. Altura 4.2 Peso 3.;
```

En el cuadro anterior se observa que:

- La variable Nombre es de caracteres y tiene una longitud máxima de 9 dígitos.
- La variable Sexo es de caracteres y tiene una longitud de un dígito.
- La variable Edad es numérica y tiene una longitud máxima de 2 dígitos sin decimales.

- La variable Altura es numérica con decimales y tiene una longitud máxima de 4 dígitos. En este caso, el punto se considera como un dígito más del campo. (El punto podría no estar en el archivo y el formato sería 3.2)

- La variable Peso es numérica sin decimales y tiene una longitud máxima de 3 dígitos

PRACTICA

Se tienen los siguientes datos de ejemplares bovinos representativos de 12 fincas ganaderas. Los ejemplares corresponden a diferentes propósitos (leche, carne y doble). Los datos son los siguientes:

FINCA	PESO	RAZA	PROPOSITO
Santa Clara	300	Holstein	leche
Mercedes	350	Holstein	leche
Los llanos	500	Brown siwss	doble
El río	510	Brown swiss	doble
San José	370	Jersey	leche
San Martín	670	Cebu	carne
El paso	470	Jersey	leche
Los toros	510	Pardo suizo	doble
El Jefe	380	Santa gertrudis	doble
El Chaparral	610	Brahman	carne
Tres ases	350	Jersey	leche
San José	580	Cebu	carne

Con los datos anteriores, cree un archivo de trabajo SAS temporal e imprima los datos.

Para imprimir los datos, agregue al programa las siguientes instrucciones:

```
PROC PRINT;
RUN;
```

Controles del Apuntador

Para todas las formas del INPUT existen los siguientes controles del apuntador (cursor) :

@n Mover el apuntador a la columna n

+n Mover el apuntador n posiciones a la derecha

@ Retener el apuntador en la línea que se está leyendo

@@ Seguir leyendo valores en la misma línea y si se termina, moverse a la siguiente línea.

Los primeros dos controles son para desplazamiento dentro del registro; los últimos dos, para el desplazamiento entre registros. **Ejemplo:**

Se tienen las siguientes líneas de datos:

SITIO													TR		REP						RENDIMIENTO										HUM	
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
T	U	R	R	I	A	L	B	A					0	1		1						1	9	5	8	.	5					6	0
T	U	R	R	I	A	L	B	A					0	2		1						2	6	8	9	.	7					6	5
T	U	R	R	I	A	L	B	A					0	3		1						1	9	3	6	.	8					5	8

Para leer estos datos, la instrucción INPUT se puede escribir de la siguiente manera, utilizando el apuntador de desplazamiento en el registro:

```
INPUT Sitio $ 1-9 @14 Tr 2. +1 Rep 1. @23 Rend 6.1 +5  Hum 2.;
```

En la instrucción anterior se indica lo siguiente:

- La variable Sitio se lee como alfabética, con un formato columnar (1-9).
- El apuntador @14, hace desplazarse al cursor de lectura en el registro hasta la columna 14. De este punto se comienza a leer la variable Tr (con formato 2.) con una longitud de dos dígitos, sin decimales.
- El apuntador +1 desplaza el cursor de lectura una columna a la derecha. De esa nueva posición se comienza a leer la variable Rep (con formato 1.) con una longitud de un dígito, sin decimales.
- El apuntador @23 desplaza el cursor de lectura hasta la columna 23. De aquí se lee la variable Rend, con formato, con una longitud de seis dígitos, incluyendo un decimal.
- El apuntador +5 hace el cursor de lectura se mueva cinco espacios para leer la variable Hum, con formato, con una longitud de dos dígitos, sin decimales.

Uso del Apuntador @

Se tiene el siguiente conjunto de datos:

1	2	3	4	5	6	7	8	9	10	11	12	13
I	D	0	0	2			M		0	8		
I	A	0	0	1			H		0	3		0
X	L	0	1	0			M		1	5		
L	K	9	5	6			H		0	6		1

Este conjunto de datos tiene las siguientes características:

- El conjunto de variables es diferente dependiendo del sexo del animal.
- Las variables # montas (para machos) y # partos (para hembras), comparten las mismas columnas (10 y 11).
- Para las hembras, se midió una variable más; número de pérdidas, la cual está en la columna 13.

El programa con la instrucción INPUT para leer estos datos es:

```
DATA A;
INPUT  Ident $ 1-5 Sexo $ 8 @;
IF Sexo = 'M' THEN
     INPUT Montas;
ELSE
     INPUT Partos Perdidas;
CARDS;
Incluir el conjunto de datos
;
PROC PRINT;
RUN;;
```

En el primer INPUT se leen, sin condiciones, las variables Ident y Sexo, que son aplicables para ambos sexos.

Se utiliza el apuntador @ para retener el cursor de registro y leer las siguientes variables de acuerdo al sexo del animal. Si el Sexo es 'M', entonces se lee la variable Montas, en caso contrario, se leen las variables Partos y Perdidas. Es necesario utilizar este apuntador, ya que cada vez que SAS encuentra un INPUT, desplaza el cursor al siguiente registro. Las variables Montas, Partos y Perdidas están en el mismo registro que Ident y Sexo

Las instrucciones alternativas (IF-THEN y ELSE), que aparecen en el ejemplo se estudiarán posteriormente.

El archivo de trabajo SAS generado por las instrucciones anteriores queda de la siguiente manera:

Obs	Ident	Sexo	Montas	Partos	Perdidas
1	ID002	M	8	.	.
2	IA001	H	.	3	0
3	XL010	M	15	.	.
4	LK956	H	.	6	1

Lectura de Varias Observaciones por Línea de Datos

En los siguientes ejemplos se considera como observación al elemento del archivo SAS que contiene los valores de todas las variables medidas a un individuo. Una línea de datos puede contener los datos de una o varias observaciones del archivo externo (o datos incluidos en el programa) que se lee para generar el archivo de trabajo SAS.

Ejemplo: Se realizó un experimento donde se toman datos de volumen de copa y altura para 13 árboles. Los datos fueron:

	ARBOL												
Variables	1	2	3	4	5	6	7	8	9	10	11	12	13
Volumen	22	6	93	62	84	14	52	69	99	98	41	85	90
Altura	36	9	67	44	72	24	33	61	64	65	47	60	51

El siguiente programa SAS genera un archivo de trabajo SAS con los datos anteriores. Estos se incluyen dentro del programa y se digitan las observaciones de varios árboles en una misma línea de datos.

```
DATA Arboles;
INPUT Volumen Altura @@;
CARDS;
22 36  6  9 93 67 62
44 84 72 14 24 52 33
69 61 99 64 98 65 41
47 85 60 90 51
;
PROC PRINT;
RUN;
```

En el ejemplo anterior, el archivo SAS tendrá 13 observaciones y dos variables. El archivo queda de la siguiente manera:

Obs	Volumen	Altura
1	22	36
2	6	9
3	93	67
4	62	44
5	84	72
6	14	24
7	52	33
8	69	61
9	99	64
10	98	65
11	41	47
12	85	60
13	90	51

La instrucción INPUT, utilizando el apuntador @@, graba una observación para cada par de datos (volumen, altura) y recorre todo el registro (línea) de entrada localizando los valores. Al llegar al final de la línea, pasa a la siguiente para leer más datos y así sucesivamente hasta llegar a la última línea de datos.

Lectura de una observación en varios registros

En algunos casos, es más fácil digitar los datos de cada observación, en varios registros del archivo externo. Cuando el número de variables es grande, y los datos no caben en la pantalla, es más cómodo digitar las variables en varias líneas. En este caso se puede utilizar el ancho de la pantalla (80 columnas) como ancho máximo del registro. *Existen paquetes como Excel, que solventan mejor este manejo de datos*. Por lo tanto, cuando se tienen muchas variables, es más sencillo digitar los datos en Excel.

Las siguientes son las formas de leer varias líneas de datos para una observación SAS:

Caso 1. Cada instrucción INPUT avanza hacia la siguiente línea de datos:

```
DATA Ejemplo;
INPUT Nombre $ 1-8 Sexo $ 11;
INPUT Edad 4-5;
```

Caso 2. Se utiliza el caracter / para avanzar a la próxima línea de datos

```
DATA Ejemplo;
INPUT Nombre $1-8 Sexo $ 11 / Edad 4-5;
```

Caso 3. Se utiliza el signo # para avanzar hacia la primera columna de la enésima línea de datos.

```
DATA Ejemplo;
INPUT  #1 Nombre $ 1-8 Sexo $11 #2 Edad 4-5;
```

Ejemplo: En el siguiente programa, cada observación está contenida en dos líneas de datos. En la primera se tiene el Nombre y el Departamento, en la segunda aparece la Edad, Sexo y Salario.

```
DATA Ejemplo;
INPUT Nombre$ 1-20 Departam$ 22-35;
INPUT Edad Sexo$ Salario;
CARDS;
Juan Carlos Morales  Cómputo
35 M 1200
Ana María Contreras  Administración
25 F 750
Alexis Brenes Torres Cómputo
39 M 1100
Rocío Fernández Coto Administración
21 F 560
;
PROC PRINT;
RUN;
```

En la instrucción INPUT se debe considerar lo siguiente:

- Cada vez que aparece un INPUT en el programa, SAS lee un nuevo registro del archivo externo, excepto cuando se utiliza el apuntador @.

- La observación en el archivo SAS se graba cuando se terminan de ejecutar todas las instrucciones del paso DATA, excepto si aparece la instrucción OUTPUT. (Esta instrucción se estudiará más adelante).

Generación o modificación de nuevas variables

Las formas más comunes de agregar, generar o modificar variables al conjunto de valores de datos en un archivo SAS son:

- Por medio de fórmulas aritméticas.

- De acuerdo a una(s) característica(s) de una(s) variable(s) ya existente(s).

- Modificando una variable ya existente.

- Utilizando las diferentes funciones que dispone SAS (ver tabla en la siguiente página con algunas de las funciones más utilizadas).

Para generar una nueva variable se debe hacer lo siguiente:

- Escoger un nombre para la variable nueva. Este nombre debe ser diferente de los nombres de las variables que ya existen. Si el nombre de la variable es igual al nombre de una variable ya

definida, los valores de la última serán sustituidos por los valores de la expresión que modifican la variable.

- Escoger la fórmula para generar o modificar la variable. Codificarla como una instrucción SAS. El nombre de la nueva variable se coloca a la izquierda del signo "=". Si utiliza un nombre de variable existente, la misma se modificará y la nueva variable reemplazará la existente.

- Si es una variable que se genera o se modifica de acuerdo a una característica de otra variable, se debe programar la instrucción IF.

- Si se necesita hacer uso de una función (raíz cuadrada, logaritmo, arcoseno, etc.) se debe indicar el nombre de la función y los parámetros requeridos.

Algunas de las funciones más utilizadas en SAS

CATEGORIA	FUNCION	DESCRIPCION	EJEMPLO EN SAS
Aritméticas	ABS	Valor absoluto	Y = ABS(X);
	MAX	Valor máximo	Y = MAX(7,10,12,X);
	MIN	Valor mínimo	Y = MIN(7,10,12,X);
	SQRT	Raíz cuadrada	Y = SQRT(X+0.5);
Matemáticas	EXP	Exponencial	Y = EXP(X);
	LOG	Logaritmo natural (base e)	Y = LOG(X);
	LOG2	Logaritmo en base 2	Y = LOG2(X);
	LOG10	Logaritmo en base 10	Y = LOG10(X);
Trigonométricas	ARCOS	El arco coseno	Y = ARCOS(X);
	ARSIN	El arco seno	Y = ARSIN(X);
	ATAN	El arco tangente	Y = ATAN(X);
	SIN	El seno	Y = SIN(X);
	COS	El coseno	Y = COS(X);
Estadísticas	CV	El coeficiente de variación	Y = CV(1,3,5,1,7,6);
	MEAN	La media aritmética	Y = MEAN(1,3,5,1,7,6);
	RANGE	El rango	Y = RANGE(1,3,5,1,7,6);
	STD	La desviación estándar	Y = STD(1,3,5,1,7,6);
	SUM	La suma	Y = SUM(1,3,5,1,7,6);
	VAR	La varianza	Y = VAR(1,3,5,1,7,6);
De fecha	DATEJUL	Convierte fecha Juliana a un valor	Y = DATEJUL(X);
	JULDATE	Convierte un valor a fecha Juliana	Y = JULDATE(Fecha);
	DATE	La fecha actual	Y = DATE( );
	YEAR	El año de la fecha	Y = YEAR(Fecha);
	MONTH	El mes de la fecha	Y = MONTH(Fecha);
	DAY	El día de la fecha	Y = DAY(Fecha);
	MDY	Convierte una fecha a un valor	Y = MDY(Mes, Día, Año)
De tiempo	HMS	Las horas, minutos y segundos	Y = HMS(Horas, Minutos, Segundos);
	HOUR	Las horas de una variable de tiempo	Y = HOUR(Tiempo);
	MINUTE	Los minutos de una variable de tiempo	Y = MINUTE(Tiempo);
	SECOND	Los segundos de una variable de tiempo	Y = SECOND(Tiempo);

Ejemplo: Se tiene el siguiente programa SAS:

```
DATA Promedio;
INPUT  Carnet $ Nota1 Nota2 Nota3;
Promedio = (Nota1 + Nota2 + Nota3)/3;
IF Promedio GE 70 THEN
   Resultad = 'GANA';
ELSE
   Resultad = 'PIERDE';
CARDS;
M001 60 50 70
M002 80 70 90
M003 60 90 75
M004 90 95 93
M005 73 45 62
;
PROC PRINT;
RUN;
```

Al ejecutar el programa, el archivo de trabajo SAS contiene la siguiente información:

Obs	Carnet	Nota1	Nota2	Nota3	Promedio	Resultad
1	M001	60	50	70	60.0000	PIERDE
2	M002	80	70	90	80.0000	GANA
3	M003	60	90	75	75.0000	GANA
4	M004	90	95	93	92.6667	GANA
5	M005	73	45	62	60.0000	PIERDE

En el ejemplo anterior considere que:

- La variable Promedio se generó a partir de Nota1, Nota2, Nota3. Se aplica una expresión aritmética para obtener el promedio de las tres variables.

- La variable Resultad se generó a partir de la variable Promedio. En este caso, esta nueva variable se generó con base en una característica de la variable Promedio. Si esta variable es mayor o igual a 70, la nueva variable asume el valor "GANA". En caso contrario, (ELSE), la variable asume el valor "PIERDE".

- A cada registro del conjunto de datos se agregan las nuevas variables.

Símbolos utilizados en expresiones aritméticas

La siguiente tabla contiene los símbolos utilizados en expresiones aritméticas. El orden de los mismos, representa la jerarquía de ejecución. Para alterar este orden, se deben utilizar paréntesis.

SIMBOLO	OPERACIÓN	EJEMPLO	INSTRUCCIÓN EN SAS
**	Exponenciación	$Y=A^3$	Y=A**3;
*	Multiplicación	Y=BxC	Y=A*B;
/	División	Y=A/B	Y=A/B;
+	Suma	Y=A+B;	Y=A+B;
-	Resta	Y=A-B	Y=A-B;

La instrucción LABEL

La instrucción LABEL se puede utilizar en un paso DATA, para dar una mejor identificación de las variables. La forma de la instrucción LABEL es:

```
LABEL variable = 'etiqueta';
```

donde variable = nombre de la variable SAS
etiqueta = descripción del contenido de la variable (máximo 40 caracteres).

Cuando se utilizan procedimientos (Pasos PROC), el nombre de la variable en las salidas será sustituido por la etiqueta de la instrucción LABEL, (excepto el PROC PRINT) dando una mejor documentación a los resultados. Ejemplo:

```
DATA Ejemplo;
INPUT Rep Trat Nplanpa Rpa;
LABEL Rep='Repetición'
      Trat='Tratamiento'
    Nplapa='Número plantas parcela'
       Rpa='Rendimiento parcela';
CARDS;
1 1 56   8.5
2 1 65  12.5
1 2 48   6.3
2 2 36   5.2
1 3 74  14.8
2 3 85  16.3
;
PROC PRINT LABEL SPLIT=' ';
RUN;
```

Como se vio anteriormente, el único procedimiento que no sustituye el nombre de las variables por las etiquetas automáticamente, es el PRINT. Si desea imprimir las etiquetas, debe indicar con las instrucciones LABEL SPLIT=' '. En este caso, se utiliza el carácter blanco ' ', para separar en líneas la etiqueta. Se puede utilizar cualquier otro caracter. La salida del programa anterior se muestra a continuación:

Obs	Repetición	Tratamiento	Número plantas parcela	Rendimiento parcela
1	1	1	56	8.5
2	2	1	65	12.5
3	1	2	48	6.3
4	2	2	36	5.2
5	1	3	74	14.8
6	2	3	85	16.3

La instrucción DELETE

Esta instrucción es utilizada cuando se desean eliminar observaciones del archivo de trabajo SAS, que cumplan cierta condición. La instrucción aparece siempre acompañada de la instrucción IF. Ejemplo:

```
IF Nota >= 70 THEN  DELETE;
```

En este ejemplo, todos los casos de las observaciones que cumplan con la condición de que la Nota "es mayor o igual" a 70 serán eliminados del archivo de trabajo SAS.

La instrucción OUTPUT

La instrucción OUTPUT se utiliza cuando se desea que la observación se grabe en el archivo de trabajo SAS. La instrucción casi siempre aparece acompañada de la instrucción IF. **Ejemplos:**

```
IF Nota < 70 THEN OUTPUT;
IF Nota < 70;
```

En este ejemplo, todas los casos de las observaciones que tienen valor "menor que" 70 en la variable Nota serán grabados en el archivo de trabajo SAS. Si se tiene un IF sin THEN; el sistema lo interpreta como: THEN OUTPUT.

El siguiente programa SAS lee un conjunto de datos y genera un archivo de trabajo SAS. Se graban sólo los registros donde la nota es mayor o igual a 70.

```
DATA Notas;
INPUT Nombre $ 1-15 Nota;
IF Nota > 70 THEN OUTPUT;
CARDS;
Carlos Torres        68
José Brenes         100
María Astúa          75
Sandra Campos        96
Luis Mora            78
Elena Troyo          65
;
PROC PRINT;
RUN;
```

La salida del programa es la siguiente:

Obs	Nombre	Nota
1	José Brenes	100
2	María Astúa	75
3	Sandra Campos	96
4	Luis Mora	78

La Instrucción IF

En algunas situaciones se requiere que el Sistema SAS realice ciertas acciones para las observaciones en el conjunto de datos cuando se cumplen ciertas condiciones. Esto puede lograrse mediante el empleo de la instrucción IF en alguna de sus modalidades. El SAS dispone de los siguientes tipos de IF:

- **IF-THEN**: Sin acción, cuando la expresión es falsa. Una sola acción, si la expresión es verdadera. En este caso, si la condición se cumple, se ejecuta sólo una instrucción.

  Sintaxis:

  IF expresión **THEN** instrucción;

  Donde expresión = Cualquier expresión válida en SAS. El resultado de evaluar la expresión será falso o verdadero.
  instrucción = lo que se ejecuta si la expresión es verdadera.

  Ejemplo:

```
IF Nota > 70 THEN DELETE;
```

Esta instrucción elimina todas las observaciones del archivo SAS, cuando se cumple la condición que el valor que contiene la variable Nota sea mayor que 70.

- **IF-THEN / ELSE**: Con acciones, cuando la expresión es falsa y verdarera. El ELSE se puede utilizar únicamente cuando se tienen condiciones binarias en la variable que se está evaluando.

  Sintaxis:

  IF expresión **THEN**
  instrucción1;
  instrucción$_n$;
  ELSE
  instrucción1;
  instrucción$_n$;

  Donde expresión = Cualquier expresión válida en SAS. El resultado de evaluar la expresión será falso o verdadero

  instrucción1 = instrucción(es) que se ejecutan(n) para la parte falsa y verdadera de la expresión.

  Ejemplo:

```
IF Nota >= 70 THEN
     Resultad = 'GANA  ';
ELSE
     Resultad = 'PIERDE';
```

  Con estas instrucciones, se está generando una nueva variable de acuerdo al valor de otra. Para todas las observaciones del archivo de trabajo SAS, cuyo valor en la variable Nota sea mayor o igual que 70, el valor que asume la variable Resultad será "GANA ". En caso contrario (condición no se cumple), la variable asumirá el valor de "PIERDE". Se ejecuta una sola acción con cualquiera de los dos posibles resultados de la condición. Sin embargo, pueden incluirse más de una instrucción, de acuerdo a la situación.

Operadores lógicos y de comparación en la instrucción IF

La expresión en la instrucción IF contiene operadores lógicos y/o de comparación. Estos operadores que utiliza el SAS se presentan a continuación.

SIMBOLO	CARÁCTER	COMPARACION
<	LT	Menor que
<=	LE	Menor o igual que
>	GT	Mayor que
>=	GE	Mayor o igual que
=	EQ	Igual que
^=	NE	Diferente que

Nota: En la instrucción IF se puede utilizar el símbolo o los caracteres de los operadores.

Los operadores lógicos que se pueden utilizar en la instrucción IF son los siguientes:

OPERADOR	COMPARACION
AND	Y
OR	O
NOT	NO

Ejemplos del uso de operadores de comparación y lógicos:

```
IF Lugar = 'CATIE' AND Variedad = 'CATURRA' THEN
     Factor = 1.078;
```

En la anterior instrucción SAS, la variable Factor asumirá un valor de 1.078 para todas las observaciones del archivo de trabajo SAS que cumplan las dos condiciones, conectadas por el operador lógico AND. Si una de ellas no se cumple, la condición es falsa.

```
IF (Lugar = 'GUAYABO' OR Lugar = 'GUAPILES') AND Variedad ='ARABIGO' THEN
     Factor = 0.887;
```

En la anterior instrucción SAS, la variable factor asumirá un valor de 0.887, para todas las observaciones del archivo de trabajo SAS, que cumpla la condición de que en su variable Lugar, los valores sean 'GUAYABO' o 'GUAPILES', y además el valor de la variable variedad sea 'ARABIGO'.

Práctica

Utilice las instrucciones adecuadas del paso DATA estudiadas anteriormente, para resolver el siguiente problema:

En tres sitios forestales se evaluó el volumen de especies latifoliadas y coníferas. Con los datos de campo obtenidos se determinó un volumen aparente (m^3), el cual debe ser multiplicado por un factor de forma de acuerdo a la especie y el sitio para obtener el volumen real (m^3). Las tablas de datos de campo y de factores de forma son:

SITIO	ESPECIE	VOL. APAR.
A	LATIF.	2.22
A	CONIF.	1.80
B	LATIF.	1.89
B	CONIF.	1.68
C	LATIF.	2.10
C	CONIF.	2.30

REFERENCIAS

SITIOS	ESPECIE	FACTOR
A	LATIF.	0.7
	CONIF.	0.7
B	LATIF.	0.6
	CONIF.	0.5
C	LATIF.	0.68
	CONIF.	0.73

Hacer lo siguiente:

a. **Crear un archivo de texto con los datos anteriores**
b. **Crear un archivo de trabajo SAS temporal que lea el archivo ASCII creado.**
c. **Generar una nueva variable que contenga el volumen real a partir del volumen aparente y el factor de forma.**
d. **Imprimir los datos.**

Los ciclos DO

Cuando se requiere ejecutar una o algunas instrucciones repetitivamente (ciclos), SAS dispone de la estructura de programación DO, la cual puede adoptar las siguientes formas.

- DO
- DO OVER
- DO WHILE
- DO UNTIL

El número de veces que se ejecuta el ciclo DO, depende de la forma de utilizar la instrucción. Todo ciclo DO termina con la instrucción END.

Ciclos controlados (DO iterativo)

Se puede ejecutar un ciclo controlado, es decir, se indica el inicio y final del ciclo. Para este tipo de ciclo se requiere de una variable de control. La forma de la instrucción es:

```
DO I = Inicio TO Final BY Incremento;
```

Donde:

I = Variable de control
Inicio = Valor inicial de la variable de control
Final = Valor final de la variable de control
Incremento = Unidades en que se incrementa el valor de la variable de control (opcional)

Ejemplo: Se tiene el siguiente conjunto de datos:

TEMPERATURAS						
Número Planta	Alta Variedad			Baja Variedad		
	1	2	3	1	2	3
1	3.5	2.5	3.0	8.5	6.5	7.0
2	4.0	4.5	3.0	6.0	7.0	7.5
3	3.0	5.5	2.5	9.0	8.0	7.0

El cuadro anterior contiene datos de altura de plantas de menta. Se tienen dos temperaturas, tres variedades por temperatura y tres plantas por variedad. Note que este conjunto de datos contiene

tres niveles jerárquicos bien definidos. El primer nivel corresponde a las temperaturas, el segundo nivel es dado por las variedades y el tercero corresponde a las plantas.

Utilizando la instrucción DO, se puede crear un archivo de trabajo SAS de estos datos, leyendo únicamente los valores para la variable altura. Las otras tres variables (Temperatura, Variedad y Planta) se pueden generar automáticamente con instrucciones DO.

El siguiente programa crea el archivo de trabajo SAS para la tabla anterior.

```
DATA Menta;
DO Temp = 'Alta', 'Baja';
   DO Varied = 1 to 3;
      DO Planta = 1 to 3;
         INPUT Altura @@;
         OUTPUT; *La instrucción OUTPUT hace que cada vez que pase por
         el ciclo, se guarde el registro en el archivo de trabajo SAS;
      END;
   END;
END;
CARDS;
3.5  4.0  3.0  2.5  4.5  5.5  3.0  3.0  2.5
8.5  6.0  9.0  6.5  7.0  8.0  7.0  7.5  7.0
;
PROC PRINT;
RUN;
```

La salida del programa anterior es la siguiente:

Obs	Temp	Varied	Planta	Altura
1	Alta	1	1	3.5
2	Alta	1	2	4.0
3	Alta	1	3	3.0
4	Alta	2	1	2.5
5	Alta	2	2	4.5
6	Alta	2	3	5.5
7	Alta	3	1	3.0
8	Alta	3	2	3.0
9	Alta	3	3	2.5
10	Baja	1	1	8.5
11	Baja	1	2	6.0
12	Baja	1	3	9.0
13	Baja	2	1	6.5
14	Baja	2	2	7.0
15	Baja	2	3	8.0
16	Baja	3	1	7.0
17	Baja	3	2	7.5
18	Baja	3	3	7.0

En el ejemplo anterior se tienen ciclos anidados. En estos casos, el ciclo más interno es el que cambia más rápido. Observando el resultado del archivo generado, se ve cómo el Número de Planta es el que varía más rápido (columna 3); el Número de Variedad (columna 2) cambia después que se ha completado un ciclo para planta y el Nivel de Temperatura (columna 1) cambia después que el ciclo de Variedad ha terminado.

Ciclos para recorrer arreglos (DO OVER)

Cuando a un conjunto de variables se les va a aplicar acciones iguales, se puede definir un arreglo (matriz unidimensional) y recorrer éste con la instrucción DO OVER, para realizar las acciones a cada una de las variables.

La forma de la instrucción DO OVER es la siguiente:

DO OVER Nombre;
    instrucción1;
        .....
    instrucción$_n$;
END;

Donde:

Nombre = Arreglo que se va a recorrer

Como se puede ver, para usar el DO OVER se requiere definir previamente un arreglo. A continuación se presenta la sintaxis de la instrucción ARRAY:

ARRAY Nombre Elementos;

Donde:

Nombre = un nombre para el arreglo, no más de ocho caracteres

Elementos = lista de los elementos (variables) que se cargan al arreglo

Ejemplo: Se quiere calcular la raíz cuadrada de los valores de 10 variables que representan conteos de diez tipos de malezas. A continuación el programa SAS:

```
DATA Malezas;
INPUT  Maleza1-Maleza10:
ARRAY Raiz Maleza1-Maleza10;
DO OVER Raiz;
       Raiz = SQRT(Raiz);
END;
CARDS;
 0  5  0 12  20  8  0  0  0 58  12 16  0  26  5 23  2  1  0  28
 7 19  5  0  18  9  0  0  8  0  19  0 10  33 16  5  8 11 14  41
 0  3  2 25   0  7  19 3  0 34
;
PROC PRINT;
FORMAT Maleza1-Maleza10 6.3;
RUN;
```

En la instrucción ARRAY se define un arreglo de nombre Raiz cuyos elementos son las diez variables que representan los conteos de las malezas. Con la instrucción DO OVER, se recorre uno a uno los elementos del arreglo y se calcula la raíz cuadrada para cada uno de ellos. Si no se utilizara un ciclo DO OVER, sería necesario escribir 10 instrucciones (una para cada variable) para transformar los datos.

Los resultados del programa anterior se muestran seguidamente:

Obs	Maleza1	Maleza2	Maleza3	Maleza4	Maleza5	Maleza6	Maleza7	Maleza8	Maleza9	Maleza10
1	0.000	2.236	0.000	3.464	4.472	2.828	0.000	0.000	0.000	7.616
2	3.464	4.000	0.000	5.099	2.236	4.796	1.414	1.000	0.000	5.292
3	2.646	4.359	2.236	0.000	4.243	3.000	0.000	0.000	2.828	0.000
4	4.359	0.000	3.162	5.745	4.000	2.236	2.828	3.317	3.742	6.403
5	0.000	1.732	1.414	5.000	0.000	2.646	4.359	1.732	0.000	5.831

Ciclos condicionados (DO WHILE y DO UNTIL)

En algunas oportunidades se requiere realizar ciclos, cuyo final está condicionado al cumplimiento de alguna condición. Los siguientes son los ciclos DO condicionados del SAS.

DO WHILE (expresión);
    Instrucción1;
    .....
    Instrucción$_n$;
END;

Donde:

Expresión = cualquier expresión válida en SAS. Esta es evaluada cada vez que se ejecuta el ciclo. Mientras la expresión sea verdadera, el ciclo se ejecuta. Cuando la expresión es falsa, el ciclo termina.

Ejemplo:

```
DATA Ciclo;
N = 0;
DO WHILE (N LE 5);
        N = N + 1;
END;
```

En el programa SAS anterior, se tiene un ciclo condicionado. Se inicializa la variable N con valor cero. La primera vez que entra al DO WHILE, la expresión se cumple (note que la expresión se cumple mientras el valor de N sea menor o igual que 5). Se entra a ejecutar las instrucciones del ciclo. En este caso, la única instrucción es incrementar el valor de N en una unidad. La expresión N = N + 1 hace que el valor de N se incremente una unidad cada vez que se ejecuta el conjunto de instrucciones correspondiente a un ciclo.

DO UNTIL (expresión);
Instrucción1;
.....
Instrucción$_n$;
END;

Donde:

Expresión = cualquier expresión válida en SAS. Esta es evaluada cada vez que se ejecuta el ciclo. Hasta que la expresión sea FALSA la instrucción se ejecuta , y la primera vez que la expresión sea verdadera, termina el ciclo.

Ejemplo:

```
DATA Ciclo;
N = 0;
DO UNTIL (N GT 5);
       N = N+1;
END;
```

En el programa SAS anterior, se tiene un ciclo condicionado. Se inicializa una variable con valor cero. La primera vez que entra al DO UNTIL la expresión no se cumple (note que la expresión se cumple cuando la variable N sea mayor que 5). Se entra a ejecutar las instrucciones del ciclo. Con los ciclos condicionados, se debe tener el cuidado de que la expresión condicionante se cumpla en determinado momento, de lo contrario se entraría en un ciclo infinito.

Las instrucciones KEEP y DROP

Estas instrucciones se utilizan cuando se va a trabajar sólo con una parte de las variables de un archivo de trabajo SAS y son útiles para economizar recursos del sistema de cómputo (disco, memoria, etc).

La instruccióon KEEP

Esta instrucción mantiene en el archivo de trabajo SAS, sólo las variables que son indicadas en la instrucción. La forma de la instrucción es la siguiente:

```
KEEP lista de variables;
```

La instrucción DROP

Esta instrucción no incluirá en el archivo de trabajo SAS aquellas variables que son indicadas en la instrucción. La forma de la instrucción es la siguiente:

```
DROP lista de variables;
```

Ejemplo: Se tiene un archivo externo que contiene 50 variables (X1,...,X50). Se van a calcular las estadísticas descriptivas sólo de algunas variables (X10...X20). En este caso no es necesario grabar en el archivo de trabajo SAS todas las variables. Se grabarán sólo aquéllas que van a ser procesadas.

Para este caso se puede utilizar el siguiente programa:

```
DATA Ejemplo; KEEP X10-X20;
INFILE  'A:\Ejemplo.dat';
INPUT X1-X50;
PROC MEANS;
RUN;
```

La instrucción RENAME

Se utiliza para cambiar nombres a variables de archivos de trabajo SAS. Puede colocarse en cualquier posición del paso Data. La forma de la instrucción es la siguiente:

```
RENAME Nombre actual = Nombre nuevo;
```

Donde:

Nombre actual = nombre actual de la variable
Nombre nuevo = nombre que sustituye al nombre actual de la variable

La instrucción TITLE

Se utiliza para añadir líneas de títulos a las salidas de los procedimientos. Se pueden colocar hasta diez líneas de títulos. La forma de la instrucción es la siguiente:

TITLEn 'Título';

Donde:

TITLEn = el título enésimo; n es el número de la línea de título
'Título' = el contenido del título enésimo

Ejemplos:

```
TITLE  'Datos agronómicos de Turrialba';
TITLE2  'Estadísticas descriptivas';
```

Estos títulos aparecerán como encabezamiento de todas las páginas que se produzcan hasta el fin de la sesión SAS, a menos que se reemplacen por otros. Si se quiere suprimir una línea de título que se ha estado imprimiendo en una sesión se emplea la instrucción TITLE sin contenido (TITLEn;).

Práctica

La siguiente tabla muestra los rendimientos en Kg por parcela de 30 parcelas de maíz. Los datos provienen de un experimento, en el cual se utilizó un diseño de parcelas divididas, con el propósito de comparar el comportamiento de 5 diferentes variedades en dos niveles de aplicación de fertilizantes. La parcela grande es la fertilización y la parcela pequeña son las variedades.

Rendimiento en Kg/parcela						
Variedades	Fertilización Alta			Fertilización Baja		
	Bloque			Bloque		
	1	2	3	1	2	3
1	7.5	9.3	6.5	2.5	1.7	3.4
2	4.0	4.5	3.0	6.0	7.0	7.5
3	3.0	5.5	2.5	1.3	2.2	2.4
4	12.3	9.8	11.4	7.2	8.7	6.5
5	6.5	8.6	6.7	4.2	3.7	3.4

Con los datos anteriores, escriba y ejecute un programa SAS que genere un archivo de trabajo SAS temporal. Las variables fertilización, variedad y bloques se deben generar automáticamente; utilizando el ciclo DO correspondiente.

En el mismo programa, genere una nueva variable que va a contener el rendimineto de maíz en kg por hectárea. El tamaño de las parcelas es el siguiente:

Para parcelas con fertilización alta, el área = 9 mts cuadrados
Para pareclas con fertilización baja, el área = 12 mts cuadrados

Imprima los datos del archivo de trabajo SAS generado.

Concatenación de archivos SAS

El sistema SAS permite generar un archivo resultante de concatenar dos o más archivos de trabajo SAS. Los archivos se pueden concatenar agregando variables (horizontalmente) mediante la instrucción MERGE o agregando registros (verticalmente) mediante la instrucción SET.

La instrucción MERGE

Esta instrucción se utiliza para concatenar dos o más archivos de trabajo SAS horizontalmente, es decir, para juntar valores de diferentes variables correspondientes a los mismos individuos.

A continuación las características más importantes del MERGE:

- Se pueden concatenar archivos, observación por observación. La primera observación de un archivo con la primera del otro, la segunda de uno con la segunda del otro, etc.
- Se pueden concatenar archivos por una o más variables de clasificación, si están ordenados por esas variables, de acuerdo a los valores de esas variables.
- El archivo resultante contendrá las variables de los diferentes archivos concatenados.
- Los archivos que se concatenan tienen que estar ordenados ya sea por posición o por las variables de clasificación.

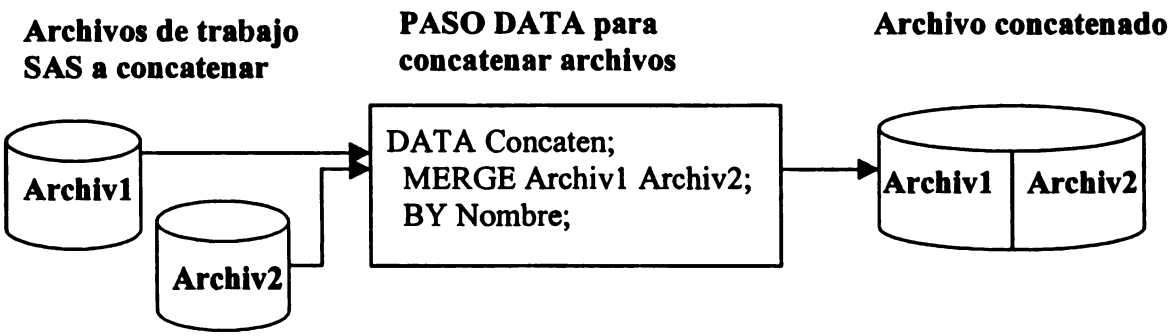
La forma de la instrucción es la siguiente:

DATA Nuevo;
MERGE Arch1...Arch$_n$;
BY Var1...Var$_n$;

Donde:

Nuevo = el archivo que contiene la concatenación
Arch1...Arch$_n$ = los archivos que se concatenan
Var1...Var$_n$ = las variables de clasificación. En este caso, todos los archivos que se concatenan deben tener estas variables.

A continuación la ilustración gráfica del MERGE:



Ejemplo: Los siguientes, son los datos de dos archivos que contienen información de maíz en diferentes sitios. En el primer archivo se tienen datos agronómicos y en el segundo datos climáticos.

Datos agronómicos de maíz para cinco sitios

Sitio	Rendimiento por Ha.	Plantas por Ha	Mazorcas por Ha
A	3572	17563	22800
D	2189	13500	16000
C	3210	16000	19600
B	2687	15632	17456
E	2741	16200	18000

Datos climáticos para cuatro sitios

Sitio	Temperatura	Precipiación	Humedad
E	26.5	3250	90
D	21.6	1890	68
B	23.4	2687	85
A	28.6	2200	82

El siguiente programa SAS, concatena verticalmente ambos archivos. Los datos se encuentran ya grabados en archivos con formato de texto. Note que los archivos no están en orden de acuerdo a la variable Sitio. También el archivo de datos climáticos no tiene los cinco sitios.

```
DATA Agronom;
INFILE 'C:\Agronom.txt';
INPUT Sitio$ Rendim Plantas Mazorcas;
PROC SORT;BY Sitio;
DATA Clima;
INFILE 'C:\Clima.txt';
INPUT Sitio$Temper Precip Humedad;
PROC SORT;BY Sitio;
DATA Total;
   MERGE Agronom Clima;
   BY Sitio;
PROC PRINT;
RUN;
```

El resultado de la ejecución del programa anterior es:

Obs	Sitio	Rendim	Plantas	Mazorcas	Temper	Precip	Humedad
1	A	3572	17563	22800	28.6	2200	82
2	B	2687	15632	17456	23.4	2687	85
3	C	3210	16000	19600	.	.	.
4	D	2189	13500	16000	21.6	1890	68
5	E	2741	16200	18000	26.5	3250	90

La instrucción SET

Esta instrucción lee las observaciones de uno o más archivos de trabajo SAS. El SET se utiliza cuando se quiere leer, hacer subconjuntos o concatenar verticalmente archivos de trabajo SAS existentes para generar un nuevo archivo. Concatenar verticalmente significa, unir valores de diferentes observaciones sobre algunas variables en común. La forma de la instrucción es la siguiente.

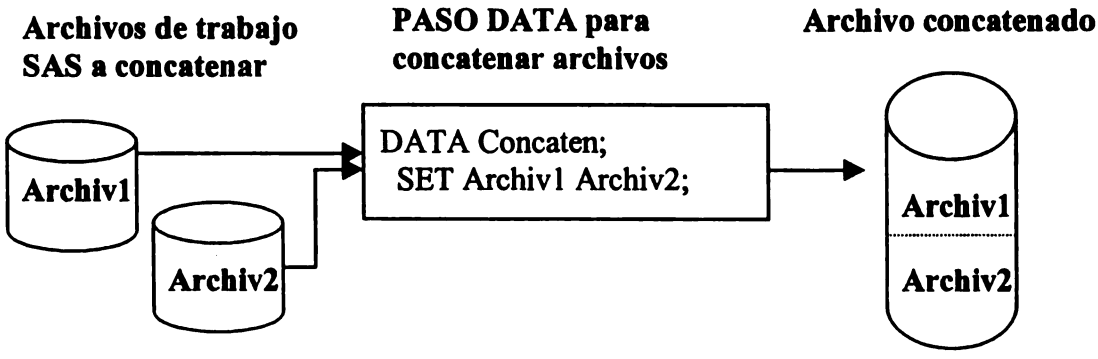
DATA Nuevo; **SET** Arch1...Arch$_n$;

Donde

Nuevo = El nuevo archivo de trabajo SAS que se genera

Arch1...Arch$_n$ = Los archivos de trabajo SAS que se unirán en el nuevo archivo

A continuación la ilustración gráfica del SET:



Ejemplo: Los siguientes son los datos de dos archivos que contienen información de maíz en diferentes sitios. En el primer archivo se tienen datos de tres sitios y en el segundo datos de cinco sitios.

Datos agronómicos de maíz para tres sitios

Sitio	Rendimiento por Ha.	Plantas por Ha	Mazorcas por Ha
A	3572	17563	22800
D	2189	13500	16000
C	3210	16000	19600

Datos agronómicos de maíz para cinco sitios

Sitio	Rendimiento por Ha.	Plantas por Ha
D	3572	17563
E	2189	13500
F	3210	16000
G	2687	15632
H	2741	16200

El siguiente programa SAS, concatena verticalmente ambos archivos. Los datos se encuentran ya grabados en archivos con formato de texto. Note que los archivos no tienen las mismas variables (el segundo no incluye datos de Mazorcas por Ha).

```
DATA Maiz1;
INFILE 'C:\Maiz1.txt';
INPUT Sitio$ Rendim Plantas Mazorcas;
DATA Maiz2;
INFILE 'C:\Maiz2.txt';
INPUT Sitio$ Rendim Plantas;
DATA Total;Set Maiz1 Maiz2;
PROC PRINT;
RUN;
```

El resultado de la ejecución del programa anterior es:

Obs	Sitio	Rendim	Plantas	Mazorcas
1	A	3572	17563	22800
2	D	2189	13500	16000
3	C	3210	16000	19600
4	D	3572	17563	.
5	E	2189	13500	.
6	F	3210	16000	.
7	G	2687	15632	.
8	H	2741	16200	.

Las variables FIRST y LAST

Se puede procesar un conjunto de datos por grupos, si se incluye la instrucción BY después de la instrucción SET o MERGE en el paso DATA. Para esto, los datos deben estar ordenados por la(s) variable(s) que identifica(n) al grupo.

Cuando se procesan archivos por grupos, se puede identificar cual es la primera y la última observación de cada grupo existente en el archivo, haciendo uso de ciertas variables automáticamente creadas por SAS, que se llaman FIRST y LAST.

La variable FIRST

Esta variable identifica cuál es la primera observación de cada grupo existente en un archivo de trabajo SAS, tomando el valor falso si no es la primera observación y verdadero si es la primera observación de un determinado grupo.

La forma de aplicar una instrucción a la primera observación de cada grupo es:

```
IF FIRST.Variable THEN Instrucciones;
```

Donde:

Variable = Variable que identifica al grupo
Instrucciones = Lo que se ejecuta si el FIRST es verdadero

La variable LAST

Esta variable identifica cuál es la última observación de cada grupo existente en un archivo de trabajo SAS. Toma valor falso, si no es la última observación y verdadero si es la última observación de un determinado grupo. La forma de aplicar una instrucción a la última observación de cada grupo es:

```
IF LAST.Variable THEN Instrucciones;
```

Donde:

Variable = Variable que identifica al grupo
Instrucciones = Lo que se ejecuta si el LAST es verdadero

Ejemplo: el siguiente programa lee un conjunto de datos (partos en vacas), lo ordena por las variables identificación de la vaca y número de parto. Crea un archivo de trabajo SAS que contiene los datos referentes sólo al último parto de cada animal.

```
DATA Partos;
INPUT Ident$ 1-5 Parto  Sexocria$  Pesocria;
CARDS;
XL555  1 M  40.3
XL555  2 M  35.6
AM001  1 H  35.2
AM001  2 M  40.6
AM001  3 H  29.5
RC210  1 H  31.6
YU115  1 M  31.8
YU115  2 M  35.7
YU115  3 H  28.6
DR570  1 H  40.1
DR570  2 M  38.7
;
PROC SORT;
    BY Ident  Parto;
DATA Ultimo; SET Partos;
BY Ident;
IF LAST.Ident THEN OUTPUT;
TITLE 'Datos de último parto';
PROC PRINT;
RUN;
```

Nota: Cuando se utiliza el BY con las instrucciones FIRST o LAST es necesario ordenar el archivo de acuerdo a las variables que identifican los diferentes grupos.

El archivo de trabajo SAS resultante al ejecutar el programa anterior queda de la siguiente forma:

Datos de último parto

Obs	Ident	Parto	Sexocria	Pesocria
1	AM001	3	H	29.5
2	DR570	2	M	38.7
3	RC210	1	H	31.6
4	XL555	2	M	35.6
5	YU115	3	H	28.6

Práctica

Para el grupo de estudiantes del curso de estadística de la promoción 1990, se tienen las calificaciones obtenidas en los quices, trabajos domiciliarios y exámenes parcial y final. La información se tiene almacenada en dos archivos ASCII. Tal como se muestra en el cuadro, el primero denominado HORIZI.DAT tiene la información sobre quices y trabajos domiciliarios y el segundo denominado HORIZ2.DAT tiene la información sobre exámenes parcial y final.

HORIZ1.DAT						HORIZ2.DAT	
Nombre	Quiz1	Quiz2	Quiz3	DOMIC1	DOMIC2	EXAMEN1	EXAMEN2
Ana	77	60	67	80	85	67	82
Carlos	80	90	87	90	93	82	90
Inés	30	50	50	70	80	60	55
José	100	95	90	100	100	100	98
Juan	75	80	77	85	90	75	80
Lorena	60	70	80	83	90	50	70
Luis	93	100	95	100	95	90	100
María	92	89	80	90	90	75	80
Mario	90	85	80	87	89	70	90
Marta	73	80	80	90	85	78	82

Para la elaboración de un informe:

a. Generar archivos SAS temporales para cada archivo ASCII
b. Ordenar cada archivo SAS por nombre
c. Crear un archivo de trabajo SAS temporal con la concatenación horizontal de los archivos SAS creados anteriormente. Para la concatenación usar la variable nombre
d. Generar la variable nota final "nota fin" de acuerdo a la siguiente ponderación:

- **promedio de quices = 20%**
- **promedio de domiciliarios = 30%**
- **examen 1 = 30%**
- **examen 2 = 20%**

e. Imprimir los datos

CAPITULO IV. Procedimientos para Salidas y Estadísticas Descriptivas

El paso DATA tiene que ver con el manejo de archivos externos y su incorporación como archivos de trabajo SAS; además, trabaja con conjuntos de datos y archivos de trabajo SAS en lo relacionado a su edición, transformación y actualización. Esto casi siempre, es una preparación de los datos para análisis por medio de los pasos PROC.

En este capítulo se explicará sobre las opciones y procedimientos para la elaboración y presentación de informes, en pantalla o impresos, resultantes del manejo y análisis de los datos. Se explica también sobre procedimientos para obtener estadísticas descriptivas.

Opciones y procedimientos para el control de salida

La mayoría de los procedimientos SAS producen "salidas" (resultados o informes que se despliegan en pantalla y que pueden imprimirse o grabarse). La forma de estas salidas pueden ser controladas por el usuario de varias maneras. Algunas de las más utilizadas se describen a continuación.

Cambiando las opciones del sistema.

Para cambiar las opciones de las salidas de los procedimientos existen dos formas:

- Se activa la ventana OPTIONS y se modifican las opciones deseadas.
- Utilizando la instrucción OPTIONS en el programa. Esta instrucción pude ir en cualquier parte del programa, pero es muy común por razones prácticas que aparezca como la primera instrucción del programa.

Las opciones más importantes a considerar en el control de salidas son:

- **LINESIZE (LS)**, determina el número de columnas por página (ancho).
- **PAGESIZE (PS)**, determina el número de líneas que se imprimen por página. Para aprovechar al máximo las líneas por página se puede emplear un tamaño de página = 58.
- **DATE**, la fecha se imprime en las salidas de los procedimientos.
- **NODATE**, la fecha no se imprima en las salidas de los procedimientos.
- **PAGENO = n**, cada vez que se ejecute el programa en la misma sesión, enumerará las páginas a partir del número indicado en n.

Ejemplo:

```
OPTIONS LS = 78   PS = 58  NODATE PAGENO=1;
```

El tamaño de Línea para aprovechar al máximo el ancho de la página es 78.

Definiendo formatos para los valores de variables.

Con el PROC FORMAT se pueden definir formatos para variables numéricas o alfanuméricas. Estos formatos se pueden asociar después dentro de otro procedimiento para cambiar los valores de las variables. La forma del PROC FORMAT es la siguiente:

```
PROC FORMAT;
     VALUE Nombre Valoractual = 'Etiqueta';
```

Donde:

Nombre = Nombre del formato (máximo 8 caracteres).
Valoractual = Valor actual que contiene la variable.
Etiqueta = Valor que sustituye al actual cuando se imprima.

En los procedimientos que se utilicen posteriormente, se asocia(n) la(s) variable(s) con el formato deseado. **Ejemplo:** se tiene el siguiente programa:

```
DATA Ejemplo;
INPUT Nombre $ Sexo Estcivil Comunid;
CARDS;
María    1 1 1
Carlos   2 2 2
José     2 3 3
Ana      1 1 2
Sonia    1 2 4
Juan     2 1 3
Mario    2 2 1
Miguel   2 3 2
;
PROC FORMAT;
     VALUE   Genero    1 ='Femenino'
                       2 =  'Masculino';
     VALUE    Civil    1 = ' Soltero'
                       2 =  'Casado'
                       3 =  'Otros';
     VALUE  Comunid 1-2 =  'Rural'
                    3-4 = ' Urbana';
PROC PRINT;
FORMAT Sexo Genero.  Estcivil Civil. Comunid Comunid.;
RUN;
```

Del programa anterior note que:

- Se definió un formato para cada variable codificada (Sexo, Estcivil, Comunidad). El nombre del formato debe tener como máximo 8 caracteres y deben ser sólo letras.
- Al PROC PRINT se le agrega la instrucción FORMAT, donde se asocian los formatos para cada variable a la cual se le definió un formato. El nombre del formato termina con un punto "." de modo que si el nombre del formato es igual al nombre de la variable no habrá confusión.

Al ejecutar el programa anterior se obtiene el siguiente resultado, donde los códigos han sido reemplazados, para su presentación, por los nombres más descriptivos de los valores de las variables.

Obs	Nombre	Sexo	Estcivil	Comunid
1	María	Femenino	Soltero	Rural
2	Carlos	Masculino	Casado	Rural
3	José	Masculino	Otros	Urbana
4	Ana	Femenino	Soltero	Rural
5	Sonia	Femenino	Casado	Urbana
6	Juan	Masculino	Soltero	Urbana
7	Mario	Masculino	Casado	Rural
8	Miguel	Masculino	Otros	Rural

Procedimientos para reorganizar archivos de trabajo SAS

Existen procedimientos para reorganizar archivos de trabajo SAS, de acuerdo a una o varias variables clasificatorias. También es posible transponer archivos de trabajo SAS, es decir, que las variables sean las observaciones y las observaciones las variables. A continuación los procedimientos.

PROC SORT

Este procedimiento se utiliza para ordenar las observaciones de archivos de trabajo SAS de acuerdo a una o varias variables. Con este procedimiento se debe utilizar la instrucción BY, seguido por el nombre de la(s) variable(s) por la(s) que se va a ordenar. La forma del procedimiento es:

```
PROC SORT Opciones ;
    BY Opción Variable Opción Variable...;
```

Las opciones que se pueden utilizar en el PROC SORT son:

- **DATA** = nombre del archivo de trabajo SAS que se va a ordenar. Si se omite esta opción, el procedimiento se aplica al último archivo creado por el programa.
- **OUT** = nombre del archivo de trabajo SAS ordenado. Si se omite esta opción, el archivo de trabajo SAS será el mismo que se está ordenando.
- **NODUPKEY** evita que en el nuevo archivo se incluyan registros para los cuales se repiten los valores de las variables por las cuales se ordena.

Las opciones de la instrucción BY son:

- **DESCENDING,** cuando se quiere ordenar una variable en forma descendente. Por defecto el ordenamiento es ascendente, por lo cual no es necesario especificarlo.

Ejemplo: a continuación se presenta un programa que ordena un conjunto de datos y crea un archivo de trabajo SAS.

```
DATA Ordena;
INPUT Sitio$  Rep Trat Rendim;
CARDS;
B 3 2 1800.65
B 1 2 2587.00
B 2 2 1745.88
A 1 2 2800.65
A 2 2 1587.00
A 3 2 3458.00
B 3 1 2233.41
B 2 1 3058.25
B 1 1 2956.50
;
PROC SORT;BY Sitio Rep Trat;
PROC PRINT;
RUN;
```

La siguiente es la salida obtenida por el programa anterior:

Obs	Sitio	Rep	Trat	Rendim
1	A	1	2	2800.65
2	A	2	2	1587.00
3	A	3	2	3458.00
4	B	1	1	2956.50
5	B	1	2	2587.00
6	B	2	1	3058.25
7	B	2	2	1745.88
8	B	3	1	2233.41
9	B	3	2	1800.65

PROC TRANSPOSE

Este procedimiento se utiliza para transponer archivos de trabajo SAS, cambiando las observaciones por variables y viceversa. Si no se utiliza la instrucción BY en el procedimiento, la transposición será para todo el archivo. Si se desea transponer por grupos dentro de un archivo SAS, se utiliza la instrucción BY.

Ejemplo: se tiene el siguiente programa SAS que traspone el archivo por una variable de clasificación (TRAT).

```
DATA Ejemplo;
INPUT Trat Rep1 Rep2 Rep3;
CARDS;
1  10.2  8.9  12.5
2   6.3  4.5   5.2
3  11.2 13.4  12.7
;
PROC TRANSPOSE OUT = Sale PREFIX = Rend;
   VAR Rep1-Rep3;
   BY Trat;
PROC PRINT;
RUN;
```

El programa anterior da los siguientes resultados:

Obs	Trat	_NAME_	Rend1
1	1	Rep1	10.2
2	1	Rep2	8.9
3	1	Rep3	12.5
4	2	Rep1	6.3
5	2	Rep2	4.5
6	2	Rep3	5.2
7	3	Rep1	11.2
8	3	Rep2	13.4
9	3	Rep3	12.7

En el programa anterior note que:

- El archivo SAS se transpone por una variable que identifica un grupo (en este caso Trat).
- Las variables Rep1, Rep2 y Rep3 pasan a ser observaciones para cada grupo identificado en la variable Trat.
- Se utiliza la instrucción OUT= Sale. Con esta instrucción se genera un nuevo archivo SAS (en este caso Sale) el cual contiene la transposición.
- Se utiliza la instrucción PREFIX para darle un nombre a la nueva variable que se genera. Si no se utiliza esta instrucción, el SAS dará el prefijo COL a las variables.

Procedimiento para listar archivos SAS

El procedimiento PRINT imprime el contenido de archivos de trabajo SAS. Se puede utilizar la instrucción BY para listar el archivo por grupos (de acuerdo a una o algunas variables de clasificación).

La forma del procedimiento es:

```
PROC PRINT;
    VAR V1..Vn;
    BY G1..Gn;
```

Donde:

V1..Vn = las variables que se desea imprimir. Si no se utiliza la instrucción VAR, el procedimiento imprime todas las variables que tiene el archivo.

G1..Gn = las variables que identifican los diferentes grupos por los que se quiere imprimir el archivo. Si no se utiliza la instrucción BY el listado será uno solo para todo el archivo.

Si se ha utilizado la instrucción LABEL y se desea que el procedimiento imprima las etiquetas, la instrucción sería de la siguiente manera:

```
PROC PRINT LABEL;
```

Procedimientos para obtener gráficas

El SAS cuenta con un módulo para la elaboración de gráficas de alta calidad. Sin embargo, su empleo adecuado requiere el manejo de varias instrucciones y muchas opciones que no pueden cubrirse en este manual. Aquí se explicará sobre el uso de unos procedimientos para la elaboración de gráficas que ayudan en el análisis de la información, pero que no son adecuadas para documentos en su presentación final.

PROC PLOT

Con este procedimiento se grafica una variable contra otra. Las siguientes son las características del PLOT:

- Se puede controlar la escala de las variables y el tamaño del gráfico.
- Si no se especifica el tamaño del gráfico, el procedimiento produce un gráfico cuadrado que ocupa el ancho de la página.

- Se puede controlar el símbolo utilizado. Si no se especifica, SAS utiliza letras: 'A' para una observación, 'B' para dos observaciones, etc.
- Se puede usar el valor de una variable como símbolo.
- Se puede sobreponer más de una gráfica en la misma figura.
- Se pueden hacer gráficos por variables que identifiquen grupos

La forma del procedimiento es la siguiente:

```
PROC PLOT;
     PLOT Y*X Opciones;
```

Donde:

Y = variable en el eje vertical
X = variable en el eje horizontal

Las opciones que se pueden utilizar en el PLOT son las siguientes:

- **OVERLAY** = Si se desea sobreponer más de un gráfico
- **HPOS** = El tamaño del eje horizontal (de 1 a 80)
- **VPOS** = El tamaño del eje vertical (de 1 a 60)
- **VAXIS** = Si se controla la escala en el eje vertical
- **HAXIS** = Si se controla la escala en el eje horizontal

Si se desea hacer gráficos por alguna(s) variable(s) que identifique grupos, se debe utilizar la instrucción BY. **Ejemplo:** el siguiente programa elabora un gráfico para dos variables:

```
DATA  Grafico;
INPUT Mes$ Ingreso @@;
CARDS;
1  1800  2   2700 3  1300  4   1400 5  2500 6   3300
;
PROC PLOT;
     PLOT Ingreso*Mes/VAXIS=1000 TO 3500 BY 500 VPOS=10 HPOS=30;
RUN;
```

La salida del programa anterior es la siguiente:

```
Plot of Ingreso*Mes.  Legend: A = 1 obs, B = 2 obs, etc.

          Ingreso |
                  |
                  |
                  |
             3500 |                              A
             3000 |
             2500 |       A                A
             2000 |  A
             1500 |             A     A
             1000 |
                  |
                  -------------------------------
                    1     2     3     4     5     6

                                  Mes
```

PROC CHART

Este procedimiento produce diagramas de barras, histogramas, diagramas de bloques y diagramas circulares. El tipo de gráfica se define con las siguientes instrucciones:

- **VBAR** histograma vertical
- **HBAR** histograma horizontal
- **BLOCK** diagrama de bloques
- **PIE** diagrama circular

El tipo de estadística se define con la opción TYPE = Las siguientes son las estadísticas disponibles:

- **FREQ** Frecuencia
- **PCT** Porcentajes
- **CFREQ** Frecuencias acumuladas
- **CPCT** Porcentajes acumulados
- **SUM** Totales
- **MEAN** Promedios

La forma de agrupar los valores se define según el tipo de variable graficada, sea numérica o de caracteres, y por opciones:

- **DISCRETE** Para agrupamiento categórico (variables discretas)
- **GROUP =** Para agrupar según otra variable
- **MIDPOINTS** = Para definir puntos medios de intervalos en variables continuas

Ejemplo: Graficar un diagrama vertical, horizontal y de bloque para agruparla según variable Com (comunidad), utilizando porcentajes:

```
PROC CHART;  (Produce un diagrama vertical)
     VBAR Ocup/DISCRETE TYPE=PCT GROUP=Com;

PROC CHART; (Produce un diagrama horizontal)
     HBAR Ocup/DISCRETE TYPE=PCT GROUP=Com;

PROC CHART; (Produce un diagrama de bloques)
     BLOCK Ocup/DISCRETE TYPE=PCT GROUP=Com;
```

Práctica

EL siguiente cuadro muestra los datos promedios mensuales de temperatura y precipitación pluvial de la década de 1981-1990 de la Ciudad de San Clemente.

MES	TEMPERATURA	PRECIPITACION
A (ENE)	20.2	800
B (FEB)	22.0	750
C (MAR)	23.5	600
D (ABR)	24.0	550
E (MAY)	25.3	650
F (JUN)	24.8	800
G (JUL)	26.0	850
H (AGO)	24.2	1000
I (SEP)	26.0	1125
J (OCT)	25.4	1400
K (NOV)	25.0	1350
L (DIC)	22.0	1000

Utilizando los datos anteriores, realice lo siguiente:

a. **Crear un archivo SAS temporal**
b. **Imprimir los datos**
c. **Hacer un gráfico de la precipitación vs. Mes**
d. **Hacer un gráfico de la temperatura vs. Mes**

Procedimientos para estadísticas descriptivas

Existen en SAS una serie de procedimientos para obtener estadísticas descriptivas. Los siguientes son algunos de los procedimientos más utilizados:

PROC MEANS

Este procedimiento produce estadísticas descriptivas univariadas simples para variables numéricas. Las siguientes son las características más relevantes del MEANS:

- Se pueden seleccionar las estadísticas a calcular.

- Util para calcular estadísticas descriptivas por grupos: utilizando la instrucción BY, el MEANS calculará las estadísticas de acuerdo a la(s) variable(s) del BY (el archivo debe estar ordenado por esta(s) variable(s)).

- La impresión de las salidas es opcional.

- Se puede crear un archivo de trabajo SAS de las estadísticas solicitadas, para ser analizadas por otros procedimientos.

Forma básica del MEANS:

```
PROC MEANS opciones;
    VAR  V1...Vn;
```

Las opciones que tiene el MEANS son las siguientes:

- **MAXDEC=n** Utiliza n decimales en las salidas

- **NOPRINT** No imprime las estadísticas. Esta opción se utiliza cuando los resultados van a ser cargados a un archivo de trabajo SAS

Las instrucciones que se pueden utilizar en el PROC MEANS son las siguientes:

- **BY** variables ; la lista de variables para calcular las estadísticas por grupos.

- **OUTPUT OUT=** Archivo ; el nombre del archivo de trabajo SAS que tendrá las estadísticas calculadas (opcional).

Las estadísticas que calcula el procedimiento se presentan en la siguiente tabla:

ESTADISTICA	DESCRIPCION
N	Número de observaciones
NMISS	Número de observaciones con valores faltantes
MEAN	La media aritmética
STD	La desviación estándar
MIN	El valor mínimo
MAX	El valor máximo
RANGE	El rango
SUM	La suma
VAR	La varianza
USS	La suma de cuadrados no corregida
CV	El coeficiente de variación
SKEWNESS	El valor de Skewness
KURTOSIS	El valor de Kurtosis
T	El valor de T
CSS	La suma de cuadrados corregida
STDERR	El error estándar

Ejemplo: el siguiente programa SAS lee un conjunto de datos y calcula la media, la desviación estándar y la suma de la variable rendimiento. Estas estadísticas se calculan por la variable Nitrógeno y los resultados se graban en un archivo de trabajo SAS llamado Estadis.

```
DATA Ejemplo;
INPUT Nitrogen Rendim @@;
CARDS;
  0   8.5   100   10.2     0   6.9   200  14.3  200  15.5
100   9.3   300   19.5   300  21.3   100   8.5    0   4.3
400  12.5   400   11.3   400  10.5
;
PROC SORT; BY Nitrogen;
PROC MEANS MEAN STD SUM NOPRINT;
     BY Nitrogen; VAR Rendim;
     OUTPUT OUT = Estadis  MEAN = MedRend STD = StdRend SUM = SumRend;
PROC PRINT;VAR Nitrogen  MedRend  StdRend  SumRend;
RUN;
```

Al ejecutar el programa, el archivo de trabajo SAS queda de la siguiente manera:

Obs	Nitrogen	MedRend	StdRend	Sum Rend
1	0	6.5667	2.11975	19.7
2	100	9.3333	0.85049	28.0
3	200	14.9000	0.84853	29.8
4	300	20.4000	1.27279	40.8
5	400	11.4333	1.00664	34.3

PROC UNIVARIATE

El procedimiento UNIVARIATE es similar el MEANS. Da las mismas estadísticas que éste. Sin embargo, da otra información como percentiles, mediana, moda. Además permite hacer una prueba de normalidad a las variables. Si se desea, también se pueden obtener gráficos que permiten analizar la forma de la distribución de los datos, como por ejemplo, el gráfico de tallo/hoja, un histograma con los datos, el gráfico de normalidad, etc. La forma del procedimiento es:

```
PROC UNIVARIATE opciones;
     VAR  V1...Vn;
```

Las opciones que se pueden incluir son:

- **NORMAL** Da varios estadísticos, para la prueba de hipótesis de que los datos tienen una distribución normal. Entre estos están : Shaphiro, Kolmogorov-Smirnov, entre otros.
- **PLOT** Realiza el gráfico de tallo/hoja, box plot y un gráfico de la probabilidad de normalidad.
- **FREQ** Realiza una tabla de frecuencias con frecuencias absolutas, relativas y acumuladas de los datos
- **NOPRINT** No imprime las estadísticas. Esta opción se utiliza cuando los resultados van a ser grabados en un archivo de trabajo SAS

Las instrucciones que se pueden utilizar en el PROC UNIVARIATE son las siguientes:

- **BY** variables ; la lista de variables para calcular las estadísticas por grupos.
- **OUTPUT OUT=** Archivo ; el nombre del archivo de trabajo SAS que tendrá las estadísticas calculadas (opcional).

Ejemplo: el siguiente programa lee un conjunto de datos, calcula estadísticas descriptivas y se realiza la prueba de normalidad.

```
DATA Normal;
INPUT Carne Nota @@;
CARDS;
 1 80 2 83 3 75 4 70 5 60 6 93 7 74 8 69
 9 72 10 86 11 82 12 95 13 55 14 85 15 77
;
PROC univariate normal;
     var Nota;
RUN
```

A continuación se muestran las salidas básicas del procedimiento y la interpretación de la prueba de normalidad. El procedimiento da más información, sin embargo, por razones prácticas de este manual, sólo se muestran las más relevantes.

The UNIVARIATE Procedure
Variable: Nota

Moments

N	15	Sum Weights	15
Mean	77.0666667	Sum Observations	1156
Std Deviation	11.0806567	Variance	122.780952
Skewness	-0.3260851	Kurtosis	-0.0548278
Uncorrected SS	90808	Corrected SS	1718.93333
Coeff Variation	14.3780147	Std Error Mean	2.86101325

Basic Statistical Measures

Location		Variability	
Mean	77.06667	Std Deviation	11.08066
Median	77.00000	Variance	122.78095
Mode	.	Range	40.00000
		Interquartile Range	15.00000

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.977108	Pr < W	0.9459
Kolmogorov-Smirnov	D	0.099975	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.019403	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.15886	Pr > A-Sq	>0.2500

Para todos los procedimientos en SAS que plantean hipótesis, la manera de proceder para aceptar o rechazarla es la siguiente:

- Definir el alfa con el que se va a trabajar
- Comparar el alfa, con la probabilidad estimada por SAS , asociada al estadístico, si:
  - Probabilidad > Alfa, entonces se acepta H_0
  - Probabildiad < Alfa, entonces se rechaza H_0
- El SAS siempre indica la Hipótesis nula que se plantea

Siempre que el SAS presente un valor de probabilidad P, esto indica que se está probando una hipótesis estadística. El usuario en este caso, siempre debe verificar cuál es la hipótesis nula (H_0) que se plantea el SAS.

La hipótesis nula que se plantea el SAS con respecto a **D** es: **H_0: D es normal**

Con las indicaciones anteriores y analizando la prueba de normalidad de Kolmogorov-Smirov estimada por el programa anterior y considerando un alfa del 5% (0.05), se concluye que los datos presentan una distribución normal, ya que Pr > **D (0.1500)**.

PROC FREQ

El procedimiento FREQ produce tablas de frecuencias en una o más dimensiones. Las siguientes son las características del procedimiento.

- Produce frecuencias absolutas y relativas, simples y acumuladas.
- Calcula medidas de asociación y pruebas estadísticas para tablas de dos dimensiones o entradas.
- Para tablas de dos entradas se puede realizar la prueba de independencia chi-cuadrado.
- Produce salidas opcionales a un archivo de trabajo SAS.

Forma del procedimiento:

```
PROC FREQ;
   TABLES Variables/OPCIONES;
          *(Si desea una tabla de dos entradas, se utiliza un asterisco);
   WEIGHT  Variables de ponderación ;
   BY  Grupos;
```

Las siguientes son las opciones del procedimiento FREQ:

EXPECTED	NOROW
DEVIATION	NOCOL
CELLCHI2	CUMCOL
CHISQ	LIST
ALL	NOCUM
NOPRINT	NOFREQ
MISSING	NOPERCENT
SPARSE	

Ejemplo: Se tiene el siguiente programa:

```
DATA Ejemplo;
INPUT Sexo Niveledu @@;
CARDS;
1 1 2 5 1 3 2 4 1 2 2 5 1 2 2 3 1 5 2 3
1 1 2 1 1 1 2 2 1 3 2 4 1 3 2 4 1 3 2 4
;
PROC FORMAT;
       VALUE Sexo      1 = 'Femenino'
                       2 = 'Masculino';
       VALUE Niveledu  1 = 'Sin escolaridad'
                       2 = 'Primaria'
                       3 = 'Secundaria'
                       4 = 'Universitaria'
                       5 = 'Otros';
PROC FREQ;
     TABLES Sexo Niveledu;
     FORMAT Sexo Sexo. Niveledu Niveledu.;
     TABLES Sexo * Niveledu / CHISQ;
     FORMAT Sexo Sexo. Niveledu Niveledu.;
RUN;
```

En el programa se utiliza la instrucción **CHISQ**, la cual es una opción del FREQ para obtener estadísticos de asociación entre las variables. La siguiente es la salida que produce el programa:

The FREQ Procedure

Sexo	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Femenino	10	50.00	10	50.00
Masculino	10	50.00	20	100.00

Niveledu	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Sin escolaridad	4	20.00	4	20.00
Primaria	3	15.00	7	35.00
Secundaria	6	30.00	13	65.00
Universitaria	4	20.00	17	85.00
Otros	3	15.00	20	100.00

Table of Sexo by Niveledu

Sexo	Niveledu					
Frequency Percent Row Pct Col Pct	Sin escolaridad	Primaria	Secundaria	Universitaria	Otros	Total
Femenino	3 15.00 30.00 75.00	2 10.00 20.00 66.67	4 20.00 40.00 66.67	0 0.00 0.00 0.00	1 5.00 10.00 33.33	10 50.00
Masculino	1 5.00 10.00 25.00	1 5.00 10.00 33.33	2 10.00 20.00 33.33	4 20.00 40.00 100.00	2 10.00 20.00 66.67	10 50.00
Total	4 20.00	3 15.00	6 30.00	4 20.00	3 15.00	20 100.00

The FREQ Procedure

Statistics for Table of Sexo by Niveledu

Statistic	DF	Value	Prob
Chi-Square	4	6.3333	0.1756
Likelihood Ratio Chi-Square	4	7.9509	0.0934
Mantel-Haenszel Chi-Square	1	3.2890	0.0697
Phi Coefficient		0.5627	
Contingency Coefficient		0.4904	
Cramer's V		0.5627	

WARNING: 100% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Sample Size = 20

Como se puede ver de la salida anterior, cuando se hacen frecuencias de dos entradas, en cada celda se tienen cuatro datos:

- **Frequency**: es la frecuencia absoluta con respecto a todos los datos
- **Percent**: es la frecuencia relativa con respecto a todos los datos
- **Row Pct**: es el porcentaje con respecto a la fila (en este caso, sexo)
- **Col Pct**: es el porcentaje con respecto a la columna (en este caso, nivel educativo)

Los estadísticos anteriores permiten analizar el cruce de las variables desde diferentes perspectivas: del total, con respecto a fila y con respecto a columna.

Al incluir la opción CHISQ, se obtienen los estadísticos que dan la probabilidad de independencia de las variables, estos estadísticos se utilizan de acuerdo a la naturaleza de las variables, por ejemplo, si son biológicas, agronómicas, económicas, sociales, etc. Para el ejemplo, el estadístico que interesa es el Chi-Square. La hipótesis que se plantea el SAS es la siguiente: **H_0: las variables son independientes.**

La probabilidad asociada a Chi-cuadrado es **0.1756**, por lo tanto, con un alfa de 0.05, se acepta H_0, es decir, existe independencia entre sexo y nivel de escolaridad.

El SAS siempre da una advertencia para estas pruebas, y es con respecto al porcentaje de celdas con valores esperados menores a cinco. Se dice que para que la prueba sea válida, las celdas con valores esperados menores a cinco, no deben ser mayores al 20% del total de las celdas. Sin embargo, esto siempre va a quedar a criterio del investigador.

Práctica

Con el propósito de evaluar la ganancia de peso en cabritos se probaron dos dietas diferentes. Los resultados se presentan a continuación:

Dieta 1	Dieta2
1.2	0.8
0.8	1.1
1.3	0.7
0.7	0.6
2.0	0.8
1.7	1.2
1.4	0.9
1.8	0.7
	0.4
	0.7

Verifique utilizando el SAS, si los datos de las ganacias de peso se distribuyen normalmente. Considere para la prueba de hipótesis un alpha del 1% y un alpha del 5%.

CAPITULO V. Algunos Procedimientos Para Análisis Estadístico de Datos Experimentales

En este capítulo se estudian algunos de los procedimientos estadísticos más utilizados, en el análisis de datos experimentales, con ejemplos del campo agropecuario y forestal. Se explica el uso del procedimiento, la manera de programarlo en SAS y la interpretación de los resultados que dan los procedimientos. También para algunos análisis, se indican los supuestos que se deben cumplir para su aplicación.

Correlación

El procedimiento CORR calcula el coeficiente de correlación para un grupo de variables. La forma del procedimiento es la siguiente:

```
PROC CORR opciones;
   VAR V1-Vn;
```

En opciones se puede indicar el tipo de coeficiente deseado. Algunos de los coeficientes disponibles en SAS son **PEARSON, SPEARMAN y KENDALL**. Si no se indica, el SAS calcula el de PEARSON.

Con el CORR se puede utilizar la siguiente instrucción:

- **WITH**, cuando se desea especificar las combinaciones de las variables las que se les desea calcular el coeficiente de correlación. **Ejemplo: PROC CORR**; **VAR** Y1 Y2; **WITH** X1-X5; En este caso, se calculan los coeficientes de Y1 con X1 a X5 e igual manera para Y2.

Ejemplo: el científico a veces está interesado en la relación que existe entre dos o más variables. Se quiere saber la interdependencia que existe entre la edad de un animal y su peso. A continuación los datos observados:

Edad (años)	1	3	2	4	2	3	2	1	2	3	4	5
Peso (kgs)	20	40	32	48	21	39	34	17	31	37	51	55

El programa SAS sería:

```
DATA Pesos;
INPUT Edad Peso @@;
CARDS;
1 20 3 40 4 48  2  21 3  39  2 34 1  17  2  31  4  51  5  55 2  32  3  37
;
PROC CORR; VAR Edad Peso;
RUN;
```

La salida obtenida por el programa anterior es la siguiente:

```
                          The CORR Procedure

                        2  Variables:    Edad      Peso

                             Simple Statistics

Variable    N      Mean      Std Dev          Sum       Minimum       Maximum
Edad       12   2.66667      1.23091     32.00000       1.00000       5.00000
Peso       12  35.41667     12.20625    425.00000      17.00000      55.00000

                  Pearson Correlation Coefficients, N = 12
                        Prob > |r| under H0: Rho=0

                                  Edad          Peso

                     Edad       1.00000       0.96003(1)
                                               <.0001(2)

                     Peso       0.96003       1.00000
                                 <.0001
```

Como se puede ver, SAS produce una matriz de correlación con el coeficiente de correlación **(1)** entre EDAD y PESO de 0.96003. El valor que se encuentra debajo del coeficiente de correlación es la probabilidad **(2)** de error cuando se rechaza la hipótesis de que el coeficiente de correlación es cero. En el ejemplo, se ve que la probabilidad de que la correlación sea cero es muy pequeña (0.0001). Por esto, se rechaza que la correlación es igual a cero, con una probabilidad de error de 0.0001, o un "nivel de confianza" de 1 - 0.0001 = 0.9999.

Análisis de regresión

Existen en SAS algunos procedimientos para hacer regresiones lineales, **REG, GLM, STEPWISE, entre otros.** Se pueden ajustar modelos lineales simples (una variable independiente) o modelos lineales múltiples (más de una variable independiente). A continuación se presenta la forma del REG, ya que es el procedimiento en SAS especializado en regresión :

```
PROC REG;
    MODEL Variables dependientes =  Variables independientes;
```

Las instrucciones que se pueden utilizar con REG son:

- **BY** Variables; Cuando se desea calcular regresión por grupos.
- **WEIGHT** Variable; Cuando se le da peso a una variable.

Regresión lineal simple

Un técnico forestal está interesado en determinar el crecimiento en diámetro de un pino a partir del volumen de la copa. Mediante un análisis de regresión se puede averiguar si existe una relación significativa entre las dos variables y expresar, por medio de una ecuación, la relación entre el crecimiento del árbol y el volumen. Esta ecuación permite predecir cuál sería el crecimiento del árbol para determinado volumen de copa.

A continuación los datos con las mediciones:

Volumen	Crecimiento	Volumen	Crecimiento
22	0.36	69	0.61
6	0.09	99	0.64
93	0.67	98	0.65
62	0.44	41	0.47
84	0.72	85	0.60
14	0.24	90	0.51
52	0.33		

Si la relación entre las dos variables fuese lineal, el modelo que predeciría el crecimiento sería:

Crecimiento = a + b * Volumen + e

El programa SAS para ajustar este modelo de regresión, utilizando el procedimiento REG, sería el siguiente:

```
DATA Regre;
INPUT Volumen Crecim @@;
CARDS;
22 0.36 6 0.09
93 0.67 62 0.44
84 0.72 14 0.24
52 0.33 69 0.61
99 0.64 98 0.65
41 0.47 85 0.60
90 0.51
;
PROC REG;
    MODEL Crecim = Volumen;
RUN;
```

Note que es necesario escribir la palabra MODEL y luego Y = X. No deben incluirse las constantes a y b del modelo.

La salida de SAS incluye un análisis de varianza, que dice si la regresión es significativa, así como los valores calculados: a (intercepto) y b (coeficiente de regresión). La salida corespondiente al ejemplo es la siguiente:

The REG Procedure
Model: MODEL1
Dependent Variable: Crecim

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.34951	0.34951	48.93	<.0001 (1)
Error	11	0.07857	0.00714		
Corrected Total	12	0.42808			

Root MSE	0.08451	R-Square	0.8165 (2)
Dependent Mean	0.48692	Adj R-Sq	0.7998 (3)
Coeff Var	17.35643		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.16244 (4)	0.05197	3.13	0.0097 (6)
Volumen	1	0.00518 (5)	0.00073989	7.00	<.0001 (7)

Lo más relevante de esta salida es:

- Ver si la regresión es significativa, es decir si hay relación entre crecimiento y volumen. El análisis de varianza en este caso prueba la hipótesis H_0: $\beta=0$ contra H_1: $\beta\neq0$. El valor 0.0001 que se observa debajo de PROB > F **(1)**, significa que se puede rechazar la hipótesis de que $\beta=0$, con una probabilidad de error de 0.0001 o menor. Esto lleva a aceptar que el valor de β es diferente de cero y se puede representar satisfactoriamente con el valor de b calculado.

- El r^2 (R Square). El R cuadrado **(2)** indica qué proporción de la variación en el crecimiento se encontró asociada con el volumen. En este caso el r^2 fue 0.8165. Esto quiere decir que el modelo lineal explica el 81,65 % de la variación.

- El r^2 ajustado **(3)** es útil cuando se desean comparar modelos iguales, con muestras diferentes de una misma población. En este caso, el r^2 ajustado es 0.7998.

- Los valores estimados para "a" (intercepto) y para "b" (pendiente). En este caso "a" **(4)** es el valor que aparece a la par de INTERCEPT (0.16244), y el de "b" **(5)** el que aparece a la par de VOLUMEN (0.00518) en la última línea. Entonces la ecuación para predecir el crecimiento sería: **Crecimiento = 0.16244 + 0.00518(Volumen)**

- El nivel de significancia para el cual se rechazaría la hipótesis que el parámetro β (5) sea 0 (es decir, que no haya regresión). En este caso se rechaza la hipótesis con nivel de significancia 0.0001 (7) . Nótese que la prueba (7) coincide con la (1) en este caso particular.

Si se desea hacer un gráfico de los datos superponiendo la línea de regresión, simplemente se añaden las siguientes líneas al programa:

```
OUTPUT  PREDICTED = Py;
PROC PLOT;
     PLOT Crecim*Volumen= '*' Py* Volumen = '+' / OVERLAY;
```

En el gráfico los datos serán denotados con '*' y la línea de regresión (valores predichos) con '+'.

Ajuste de modelos cuadráticos y cúbicos

Algunas veces la relación entre dos variables es cuadrática o cúbica. Una relación típica de estos modelos es, por ejemplo, cuando la variable independiente (X) es un fertilizante aplicado en varias dosis. Al aumentar la dosis, aumenta el rendimiento, pero el incremento se hace cada vez menor hasta que se llega a un punto en que el rendimiento empieza a bajar (incrementos negativos) debido a que dosis altas del fertilizante producen efectos dañinos en el cultivo.

El siguiente programa hace una regresión cuadrática de rendimiento de maíz en función de dosis de fertilizantes:

```
DATA Maiz;
INPUT Dosis Rend @@;
CARDS;
  0 12.5
100 18.6
200 24.6
300 28.3
400 21.3
500 14.4
;
PROC GLM;
    MODEL Rend = Dosis Dosis*Dosis;
RUN;
```

El modelo cuadrático que se está ajustando es el siguiente:

$$\mathbf{Rend = a + b*Dosis + c*Dosis^2 + e}$$

La siguiente es la salida producida por el programa anterior:

```
                          The GLM Procedure

                      Number of observations    6

Dependent Variable: Rend

                                  Sum of
Source                     DF       Squares      Mean Square    F Value    Pr > F

Model                       2     169.4455714     84.7227857      21.45    0.0167
Error                       3      11.8494286      3.9498095

Corrected Total             5     181.2950000


                  R-Square     Coeff Var      Root MSE     Rend Mean

                  0.934640      9.961969      1.987413      19.95000


Source                     DF       Type I SS     Mean Square    F Value    Pr > F

Dosis                       1       6.4812857       6.4812857       1.64    0.2902
Dosis*Dosis                 1     162.9642857     162.9642857      41.26    0.0076


Source                     DF     Type III SS     Mean Square    F Value    Pr > F

Dosis                       1     168.1555135     168.1555135      42.57    0.0073
Dosis*Dosis                 1     162.9642857     162.9642857      41.26    0.0076


                                                  Standard
Parameter                       Estimate             Error     t Value    Pr > |t|

Intercept                   11.46428571       1.80124579        6.36      0.0078
Dosis                        0.11055000       0.01694304        6.52      0.0073
Dosis*Dosis                 -0.00020893       0.00003253       -6.42      0.0076
```

La salida de SAS es similar a la mostrada en la regresión lineal simple, excepto que ahora se tiene un parámetro más, con su probabilidad de que sea significativo. Para saber si el modelo cuadrático se ajusta mejor que el lineal, se deben comparar los cuadrados medios de ambos modelos. Si la reducción es importante, entonces quiere decir que el modelo cuadrático es mejor. También se puede ver si el parámetro c es estadísticamente significativo, lo mismo que si el R cuadrado aumentó considerablemente.

El R cuadrado del modelo cuadrático siempre será más alto que el del modelo lineal (lo mismo aplica para el modelo cúbico), pero se debe considerar si este aumento en el R cuadrado es suficiente como para complicar el modelo con un factor más.

Si se desea ajustar un modelo cúbico, es decir,

$$\textbf{Crecimiento} = a + b*\text{volumen} + c*\text{volumen}^2 + d*\text{volumen}^3 + e$$

Simplemente se escribe el modelo en el programa de la siguiente manera:

```
PROC GLM;
     MODEL Rend=Dosis Dosis*Dosis Dosis*Dosis*Dosis;
```

La salida incluirá el parámetro adicional (d), así como la probabilidad de que sea significativo. Las mismas observaciones anteriores aplican aquí para saber si el modelo cúbico es mejor que el cuadrático.

Regresión Múltiple

Algunas veces se tiene un modelo de regresión, en el cual la variable depende de varias variables independientes (X). Se puede entonces ajustar un modelo de regresión múltiple.

Ejemplo: se cree que el rendimiento en kgs, de parcelas de maíz depende de la cantidad de nitrógeno aplicado al suelo, la altura promedio de las plantas (cm) y el número de mazorcas. En este caso se tendría el modelo de regresión múltiple:

$$\textbf{Rendimiento} = a + b1*N + b2*\text{Altura} + b3*\text{Mazorcas} + e$$

El siguiente programa SAS, ajusta un modelo de regresión múltiple para el ejemplo anterior:

```
DATA RegreM;
INPUT  Rend Nitrog Altura Mazorcas @@;
CARDS;
18.56 120 120 54 21.34 126 124  68 25.58 128 138  75 33.28 133 140  88
13.23 110 120 28 12.28 112 135  27 11.19 115 129  29 12.12 119 127  45
17.28 112 127 37 17.12 110 149  31 17.28 122 158  51 18.28 121 148  67
10.20  99 128 38
;
PROC GLM;
     MODEL Rend = Nitrog Altura Mazorcas;
RUN;
```

La salida del programa anterior es la siguiente:

The GLM Procedure

Dependent Variable: Rend

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	393.7073807	131.2357936	11.27	0.0021
Error	9	104.8335885	11.6481765		
Corrected Total	12	498.5409692			

R-Square	Coeff Var	Root MSE	Rend Mean
0.789719	19.48198	3.412942	17.51846

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Nitrog	1	334.4374613	334.4374613	28.71	0.0005
Altura	1	2.3874562	2.3874562	0.20	0.6615
Mazorcas	1	56.8824631	56.8824631	4.88	0.0545

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Nitrog	1	9.90510226	9.90510226	0.85	0.3805
Altura	1	1.35163720	1.35163720	0.12	0.7412
Mazorcas	1	56.88246315	56.88246315	4.88	0.0545

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-18.67628535	22.30818942	-0.84	0.4242
Nitrog	0.18955708	0.20556043	0.92	0.3805
Altura	0.02907442	0.08535126	0.34	0.7412
Mazorcas	0.20439080	0.09249135	2.21	0.0545

La salida es similar a las anteriores (regresión lineal simple y cuadrática). En este caso se tiene que:

- El valor de F de 11.27 se usa para probar la hipótesis nula de que

  H_0: b1 = b2 = b3 = 0 (o sea que no hay regresión)

- Los parámetros de este modelo son:

  Rend = -18.68 + 0.19(Nitrog) + 0.0291*Altura + 0.2044*Mazorcas

- El nivel de significancia para probar que cada parámetro es diferente de cero, o sea, si está verdaderamente contribuyendo al modelo. En este caso se puede ver que el único parámetro significativo es el número de mazorcas (0.0545). Por lo tanto, este modelo necesitaría incluir solo la variable Mazorcas, porque las otras no explican la variación en el rendimiento.

Como se puede notar en la salida, se tienen dos tipos de sumas de cuadrados. **Ver la sección 'Tipos de sumas de cuadrados calculados por el GLM'**, para una mejor explicación sobre este tema.

Regresión no lineal

Cuando a los datos es difícil ajustar un modelo lineal, el SAS dispone del procedimiento NLIN. Este procedimiento facilita ajustar modelos de regresión no lineal. A diferencia de procedimientos para ajustar regresiones lineales (GLM, REG, STEPWISE), con el NLIN se debe indicar el modelo completo que se va a ajustar, es decir, no sólo las variables del modelo, sino también los parámetros. Además, se deben indicar los valores iniciales de los parámetros.

La siguiente es la forma básica del PROC NLIN:

```
PROC NLIN BEST=Número METHOD =Método;
     MODEL Y=Expresión;
     PARAMETER Parámetros=Valores;
```

Con la opción BEST, se indica el número de las mejores sumas de cuadrados residuales que se imprimirán. Si no se especifica, el NLIN imprimirá todas las sumas de cuadrados residuales

Los métodos que dispone el SAS son : **GAUSS, MARQUARDT, NEWTON, GRADIENT y DUD**. Si el método no es especificado, el SAS aplicará el método DUD.

Otras instrucciones que se pueden utilizar en el NLIN son las siguientes:

BY variables; para aplicar la regresión por grupos de acuerdo a una o varias variables clasificatorias.

OUTPUT OUT = Archivo Estadisticas=Variable;

Las estadísticas principales que se pueden obtener son:

- **PREDICTED** ,valores predichos
- **RESIDUAL** , valores residuales
- **L95**, el límite inferior de confianza al 95%
- **U95**, el límite superior de confianza al 95%

Ejemplo: el siguiente programa, lee un conjunto de datos y ajusta el modelo exponencial negativo. Se grafican los valores observados y estimados por el modelo.

```
TITLE 'Modelo exponencial negativo: Y=B0*(1-EXP(-B1*X))';
DATA Nolineal;
INPUT X Y @@;
CARDS;
020 0.57 030 0.72
040 0.81 050 0.87
060 0.91 070 0.94
080 0.95 090 0.97
100 0.98 110 0.99
120 1.00 130 0.99
140 0.99 150 1.00
160 1.00 170 0.99
180 1.00 190 1.00
200 0.99 210 1.00
;
PROC NLIN  BEST=10 METHOD=MARQUARDT;
   PARMS B0=0 TO 2 BY 0.5
         B1=0.01 TO 0.09 BY 0.01;
   MODEL Y=B0*(1-EXP(-B1*X));
   OUTPUT OUT=Sale PREDICTED=PY;
PROC PLOT;
   PLOT Y*X='O' PY*X='P'/OVERLAY VPOS=15 HPOS=30;
RUN;
```

A continuación los resultados obtenidos por el programa anterior:

```
                 Modelo exponencial negativo: Y=B0*(1-EXP(-B1*X))              1

                            The NLIN Procedure
                               Grid Search
                           Dependent Variable Y

                                                   Sum of
                        B0            B1          Squares

                    1.0000        0.0400        , 0.00140
                    1.0000        0.0500          0.0168
                    1.0000        0.0600          0.0552
                    1.0000        0.0300          0.0666
                    1.0000        0.0700          0.0973
                    1.0000        0.0800          0.1365
                    1.0000        0.0900          0.1708
                    1.0000        0.0200          0.4193
                    1.5000        0.0100          0.9757
                    1.0000        0.0100          2.1653
```

The NLIN Procedure
Iterative Phase
Dependent Variable Y
Method: Marquardt

Iter	B0	B1	Sum of Squares
0	1.0000	0.0400	0.00140
1	0.9961	0.0419	0.000580
2	0.9962	0.0420	0.000577
3	0.9962	0.0420	0.000577
4	0.9962	0.0420	0.000577

NOTE: Convergence criterion met.

Estimation Summary

Method	Marquardt
Iterations	4
R	2.552E-7
PPC(B1)	1.028E-8
RPC(B1)	6.394E-7
Object	2.56E-10
Objective	0.000577
Observations Read	20
Observations Used	20
Observations Missing	0

NOTE: An intercept was not specified for this model.

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Regression	2	17.6717(1)	8.8359	275733	<.0001(3)
Residual	18	0.000577	0.000032		
Uncorrected Total	20	17.6723(2)			
Corrected Total	19	0.2439			

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
B0	0.9962(4)	0.00161	0.9928	0.9996
B1	0.0420(5)	0.000398	0.0411	0.0428

Approximate Correlation Matrix

	B0	B1
B0	1.0000000	-0.5558957
B1	-0.5558957	1.0000000

```
Plot of Y*X.   Symbol used is 'O'.
                    Plot of PY*X.  Symbol used is 'P'.

   Y |
 1.0 |            OOOOOOOOOOOO
     |         OOO
     |        O
     |       O
 0.8 |      O
     |
     |    O
     |
 0.6 |
     |   O
     |
     |
 0.4 |
     |
     --------|----------|-------|-------
     0        90        180       270

                          X

NOTE: 20 obs hidden.
```

Como se puede ver de la salida anterior, el PROC NLIN da una serie de estadísticos, sin embargo, se consideran sólo los más relevantes. A continuación la interpretación de la salida.

Lo primero que se debe verificar es la significancia del modelo. Se debe comparar el valor asociado a la probabilidad de F, con el alfa deseado. **Pr > F (3),** para estos datos el valor es 0.0001. Por lo tanto, se concluye que los datos se ajustan para un alfa del 5%.

Una vez verificada la significancia del modelo, se procede a calcular el coeficiente de regresión (r^2). La manera de obtenerlo es:

$$\frac{\text{suma de cuadrados de la regresión} \quad \textbf{(1) 17.6717}}{\text{suma de cuadrados total no corregida } \textbf{(2) 17.6723}}$$

Al aplicar la fórmula, se obtiene que r^2=0.9999, por lo tanto, la variabilidad de Y es explicada en un 99.99% por el modelo.

Por último, se despeja el módelo de acuerdo a los coeficientes estimados. B_0 **(4)** y B_1 **(5),** quedando de la siguiente manera: **Y=0.9962*(1-EXP(-0.0420*X)).**

En el programa también se graficaron los valores observados y estimados por el modelo. Como se puede ver, estos valores son muy similares, comprobando con esto, el excelente ajuste de los datos al modelo aplicado.

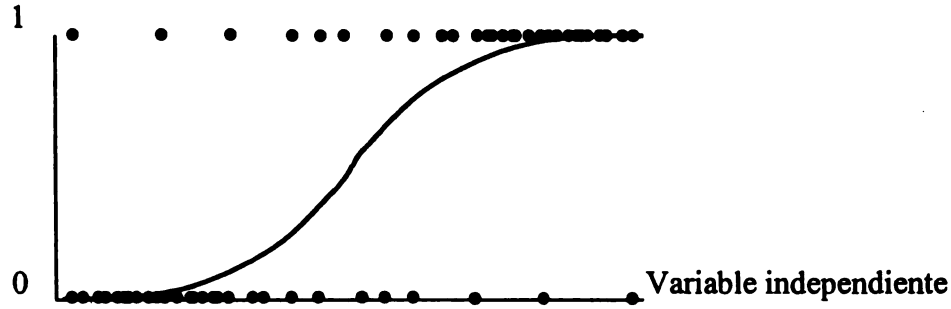
Regresión logística

La regresión logística es un método estadístico que permite modelar respuestas dicotómicas, en donde hay solamente dos posibles respuestas. En general se transforman estas dos respuestas a las cifras "0" y "1"; sin embargo, el significado puede ser, para dar algunos ejemplos, "presente" y "no presente", "vivo" y "muerto", "pertenencia a categoría X" y "pertenencia a categoría Y".

La regresión logística se aplica frecuentemente en el campo de estudios de sobrevivencia, en estudios de la salud y en la epidemiología.

Obviamente, en la regresión logística con solo dos opciones de respuesta, la "filosofía de la predicción" debe ser diferente de la de la regresión clásica. Lo que se modela es la probabilidad $P(Y=1)$ que un evento ocurra para un nivel (o rango) dado de la variable predictora. Por ejemplo, en un estudio epidemiológico de plantas, podría ser la probabilidad de que un individuo se infecte, dependiendo de la distancia de una fuente definida de infección. La "nube" de puntos también es diferente de lo que conocemos de la regresión clásica, porque la variable de respuesta exhibe solamente valores en dos niveles.

Grafico de los datos en una regresión logística



El grafico muestra la "nube" de puntos en el caso de la regresión logística, aquí solamente existen las respuestas "0" y "1". La curva ajustada nos da para cada uno de los valores de la variable independiente una estimación de la proporción de respuestas "1".

Más formalmente, la proporción estimada $\hat{p}$ que encontramos en nuestro estudio muestral para la variable de respuesta Y, la expresamos como función de la variable predictora X_1 como:

$$\hat{p} = \frac{e^{b_0+b_1X_1}}{1+e^{b_0+b_1X_1}}.$$

El anterior es el modelo logístico, que se usa en la regresión logística. Hay otros modelos, pero el modelo logístico es el más usado por tener varias características favorables.

El cálculo de estimaciones buenas y útiles de los coeficientes se hace a través de procedimientos iteractivos aplicando la técnica de la máxima verosimilitud. El método de mínimos cuadrados (que es un caso especial y simple de la técnica de la máxima verosimilitud), no se puede aplicar para llegar a estimaciones en el caso de la regresión logística porque sus suposiciones básicas no se cumplen: la respuesta dicotómica hace que la distribución de la variable de respuesta es no-normal (sino sigue obviamente la distribución binomial). Esto también hace que el nivel de significancia que se dan en las salidas de programas estadísticos no se basan en la distribución de *t*, sino se lo aproxima por la distribución chi-cuadrado.

En SAS, se puede obtener estimaciones del modelo logístico por medio del procedimiento PROC LOGISTIC cuya estructura básica se describe a continuación:

```
PROC LOGISTIC opciones;
  MODEL variable de respuesta = variables independientes/ opciones;
```

Las siguientes son algunas de las opciones de la instrucción PROC LOGISTIC:

DATA=	Aquí se indica el archivo SAS que se quiere usar; si no se especifica, SAS usará el ultimo archivo SAS creado.
COVOUT	Hace que el archivo de resultados (que se indica en OUTEST=) también contenga la matriz de variancias-covariancias.
DESCENDING	Esta opción invierte el orden de la variable de respuesta. Si simultáneamente se especifica ORDER= , primero se genera la orden según ORDER= y después se invierte este orden con DESCENDING.
INEST=	Se pueden especificar valores iniciales para el proceso iteractivo de buscar estimaciones con el método de máxima verosimilitud.
NOPRINT	Hace que el programa no imprima los resultados (en la ventana Output)
ORDER=	Especifica en que orden se quieren los datos de la variable de respuesta. Se tienen las siguientes opciones ORDER= DATA | INTERNAL | FORMATTED. Con ORDER= DATA: los niveles de la variable de respuesta se ordenan según como están en el archivo de entrada. Con ORDER= INTERNAL: se ordenan los niveles según el valor no formateado. Con ORDER= FORMATTED: se ordenan los niveles según el valor formateado. El estándar es ORDER= FORMATTED, si no se especifica un formato, el estándar es ORDER= INTERNAL.
OUTEST=	Se indica un nombre de archivo SAS que contendrá los resultados
SIMPLE	Esta opción hace que se impriman algunas estadísticas simples para cada una de las variables predictoras: media, desviación estándar, valor mínimo, valor máximo

Entre las opciones de la instrucción MODEL se tienen:

PLCL — Muestra los intervalos de confianza de los parámetros estimados basados en la razón de verosimilitudes

WALDRL — Muestra los intervalos de confianza de los parámetros estimados usando el método de Wald

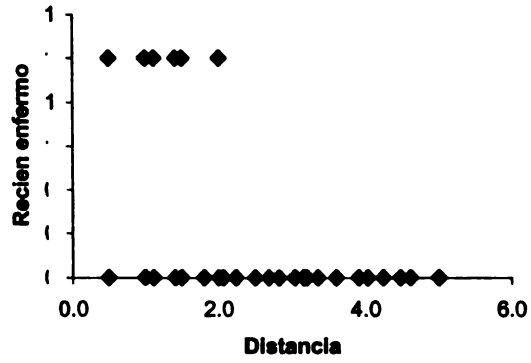
RSQUARE — Da el R^2 generalizado para medir el ajuste del modelo

SELECTION = — Aquí se indica el método de selección de variables a ser utilizado en el modelo. Los métodos son: BACKWARD, FORWARD, NONE, STEPWISE, SCORE. NONE es la opción por defecto y ajusta el modelo completo solicitado

Ejemplo: Un investigador está interesado en describir al proceso de la diseminación de una enfermedad vegetal. Él quiere probar la hipótesis que la distancia de la planta enferma más cercana influye sobre la probabilidad de que una planta se enferme. Además, tiene la hipótesis que el viento tiene un efecto sobre la diseminación. Por esto, seleccionó al azar una muestra de plantas (sanas y recientemente enfermas) y midió la distancia y la dirección de planta enferma mas cercana. Como ángulo se usa el ángulo relativo a la dirección prevalente del viento.

El conjunto de datos tiene el siguiente formato:

```
Columna 1: planta enferma ("1") o sana ("0")
Columna 2: distancia a la planta enferma más cercana
Columna 3: ángulo hacia la planta enferma más cercana
           medido como desviación de la dirección
           prevalente del viento.
 0   1.000000   1.570796
 0    .000000   0.000000
 1   0.500000   3.141593
 0   0.500000   0.000000
.....
.....
.....
 0   4.609772   0.708626
 0   5.000000   0.643501
 1   0.500000   0.000000
 0   1.000000   0.000000
 1   1.000000   3.141593
 1   0.500000   3.141593
```



En el programa que se usa, se pretende generar las estimaciones de las coeficientes de regresión, se quiere poder hacer una conclusión probabilística acerca de la significancia estadística de las mismas y se quiere calcular directamente intervalos de confianza. Para los intervalos de confianza hay que especificar opciones bajo el comando MODEL; PLCL y WALDRL. Estas opciones producen los intervalos, usando dos diferentes métodos (el método log-likelihood (razón de verosimilitudes), y el método de Wald.

El programa SAS que se usa para el análisis es el siguiente:

```
DATA tomate;
  INFILE 'c:\logreg.txt' FIRSTOBS = 6;
  INPUT enf dist angulo;
RUN;

PROC LOGISTIC DESCENDING;
 MODEL enf = dist angulo / PLCL WALDRL;
RUN;
```

Este programa produce la siguiente salida:

The LOGISTIC Procedure
Model Information (1)

Data Set	WORK.TOMATE
Response Variable	enf
Number of Response Levels	2
Number of Observations	540
Link Function	Logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	enf	Total Frequency
1	1	80
2	0	460

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	455.042	438.843
SC	459.334	451.718
-2 Log L	453.042	432.843

Testing Global Null Hypothesis: BETA=0 (2)

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	20.1989	2	<.0001
Score	13.7272	2	0.0010
Wald	13.7636	2	0.0010

Analysis of Maximum Likelihood Estimates (3)

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-0.8131	0.3379	5.7903	0.0161
dist	1	-1.2290	0.3452	12.6713	0.0004
angulo	1	0.0724	0.0918	0.6225	0.4301

Odds Ratio Estimates (4)

Effect	Point Estimate	95% Wald Confidence Limits	
dist	0.293	0.149	0.576
angulo	1.075	0.898	1.287

Association of Predicted Probabilities and Observed Responses (5)

Percent Concordant	55.9	Somers' D	0.290
Percent Discordant	26.9	Gamma	0.350
Percent Tied	17.1	Tau-a	0.073
Pairs	36800	c	0.645

Profile Likelihood Confidence

Interval for Parameters (6)

Parameter	Estimate	95% Confidence Limits	
Intercept	-0.8131	-1.4761	-0.1495
dist	-1.2290	-1.9530	-0.6041
angulo	0.0724	-0.1069	0.2539

Wald Confidence Interval for Adjusted Odds Ratios (7)

Effect	Unit	Estimate	95% Confidence Limits	
dist	1.0000	0.293	0.149	0.576
angulo	1.0000	1.075	0.898	1.287

La salida inicialmente, muestra la información general de los datos y del modelo usado (**1**). Después, presenta las pruebas del ajuste del modelo completo (**2**). Aquí, el SAS ofrece resultados de tres diferentes métodos (Likelihood Ratio, Score y Wald), normalmente son muy similares. Al analizarlos se concluye que el modelo contribuye en forma significativa a la predicción de la variable de respuesta ($p<0.0001$, $p=0.0010$ y $p=0.0010$ respectivamente). En (**3**) se observan las estimaciones de los coeficientes, con su error muestral y significancia estadística. Aquí distancia es significativo ($p=0.0004$), y ángulo no ($p=0.4201$). En (**4**) se tienen los "odds ratios", derivados de las estimaciones de los coeficientes, por ejemplo: $e^{-1.229}=0.293$. Observe que los intervalos de confianza no son simétricos. La parte (**5**) de la salida muestra el número de concordancia de pares de observaciones (columna izquierda), y 4 coeficientes de correlación de rangos (columna derecha). Los intervalos de confianza para las estimaciones de los parámetros se muestran en (**6**). Por último, en (**7**) se observan los intervalos de confianza para los odds ratios, calculados según el método de Wald, para un cambio de una unidad de cada una de las variables explicatorias. Con la instrucción UNITS (ver manual del SAS) se puede especificar este valor.

Práctica

Se realizó un experimento con el fin de evaluar la respuesta de 4 niveles de nitrógeno en el rendimiento del cultivo de sorgo. Cada nivel tuvo 5 repeticiones. El rendimiento obtenido (kg/parcela) aparece en el cuadro siguiente:

REPETICION	NIVEL DE NIT.	RENDIMIENTO (kg/parc
1	0	25.4
2	0	34.2
3	0	21.2
4	0	36.9
1	50	34.3
2	50	36.8
3	50	32.1
4	50	40.6
1	150	40.2
2	150	47.5
3	150	39.3
4	150	49.7
1	100	36.4
2	100	43.2
3	100	34.2
4	100	43.3

Resolver los siguientes aspectos:

a. **Crear un archivo de texto en la unidad "A" y nombrarlo EJEMPLO.DAT**

b. **Hacer un programa SAS que lea el archivo creado**

c. **Obtener los promedios de rendimiento para cada nivel de nitrógeno y guardarlos en un archivo de trabajo SAS temporal.**

d. **Calcular regresión simple y cuadrática de las variables rendimiento y niveles de nitrógeno, utilizando los promedios.**

e. **Elaborar un análisis gráfico de los valores observados y predichos.**

f. **Determinar cúal de los dos modelos es el que mejor explica la variabilidad del rendimiento en función del nitrógeno.**

Análisis de Medias

Existen una serie de procedimientos para comparar medias, es decir, a través de pruebas T, o bien, análisis de varianza. Con SAS, se puede programar cualquier diseño que se haya implantado en el campo, lo único que se requiere es conocer el modelo matemático del diseño y los diferentes errores que considera el diseño. A continuación, algunos procedimientos en SAS para la comparación de medias.

Comparación de dos medias muestrales

El procedimiento en SAS para comparar las medias de dos muestras independientes es el TTEST. La forma de este procedimiento es:

```
PROC TTEST;
   CLASSES Tratamiento; *(la variable tratamiento debe tener únicamente
                         dos valores);
   VAR Variables de respuesta;
```

La prueba "t" de Student proporciona el método para comparar las medias de dos muestras. Considere, por ejemplo, que se tienen datos de dos procedencias de pino hondureño, las que van a compararse sobre la base de su producción en volumen durante cierto período. Los volúmenes (en m^3) de 11 árboles de cada procedencia fueron los siguientes:

Procedencia 1: 11 5 9 8 10 11 10 8 11 8 8

Procedencia 2: 9 6 9 9 13 8 6 5 6 10 7

Para probar la hipótesis de que no hay diferencias entre las medias de las procedencias (i.e. hipótesis nula, H_0: $\mu 1 = \mu 2$), SAS realiza la prueba de "t" bajo las alternativas: a) varianzas iguales; b) varianzas diferentes. Además, SAS hace una prueba de igualdad de varianzas para que, con base en el resultado de esa prueba, el usuario escoja la prueba de "t" correspondiente.

El programa SAS para analizar estos datos sería como sigue:

```
DATA Medias;
INPUT Proced Volumen @@;
CARDS;
1  11 1  5 1  9 1  8 1 10 1 11 1 10 1  8 1 11 1   8 1  8
2   9 2  6 2  9 2  9 2 13 2  8 2  6 2  5 2  6 2  10 2  7
;
PROC TTEST;
    CLASS Proced;
    VAR Volumen;
RUN;
```

La instrucción CLASS sirve para indicar la variable que identifica los dos grupos que se van a comparar. En este ejemplo, se trata de dos procedencias que se especifican en la variable Proced. Esta variable toma valores 1 y 2 únicamente. El procedimiento TTEST hace una prueba de "t" para la(s) variable(s) indicada(s) en la instrucción VAR, usando la variable indicadora incluída en la instrucción CLASS como criterio de agrupamiento.

La salida que produce SAS es la siguiente:

The TTEST Procedure

Statistics

Variable	Class	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err
Volumen	1	11	7.7612	9	10.239	1.2884	1.8439	3.2359	0.556
Volumen	2	11	6.4389	8	9.5611	1.6237	2.3238	4.0781	0.7006
Volumen	Diff (1-2)		-0.866	1	2.8657	1.6048	2.0976	3.0291	0.8944

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
Volumen	Pooled	Equal	20	1.12	0.2768 (1)
Volumen	Satterthwaite	Unequal	19	1.12	0.2775 (2)

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
Volumen	Folded F	10	10	1.59	0.4775 (3)

Para interpretar esta salida, lo primero que se debe examinar es la prueba de varianzas (última línea). La probabilidad de obtener un valor mayor que F = 1.59 es muy alta (0.4775) **(3)**, lo que indica que no hay diferencias significativas entre las varianzas. Para poder rechazar H_0, el valor calculado debe ser menor o igual que la probabilidad de error establecida (por ejemplo, 0.05).

Al aceptar que las varianzas son iguales, se usa la prueba de "t" para varianzas iguales (equal). El valor obtenido para t es 1.12 y el valor para la probabilidad de error es de 0.2768 **(2)**. En este caso, la probabilidad de error es alta (mayor que 0.05), por lo tanto no se puede rechazar la hipótesis de igualdad de promedios. En conclusión, no existen diferencias en volumen entre las dos procedencias.

Comparación de medias de datos apareados

Considere un experimento donde se aplican dos drogas somníferas (A y B) a un grupo de individuos y se midió el número de horas extra de sueño. La droga A, se aplicó a cada individuo durante los primeros ocho días, y la droga B, los siguientes ocho días. Los siguientes datos corresponden al promedio de horas extra de sueño para cada individuo para cada droga:

Individuo	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Droga A	2	1	0	2	1	2	1.3	2	3	2	1	0
Droga B	3	2	1	3	3	4	3.8	2	4	4	3	3

Se quiere saber si la droga B produce efectos diferentes que la droga A. El método para comparar medias apareadas, se basa en calcular la diferencia dentro de cada par y hacer una prueba de t usando la varianza de las diferencias. Entre más cerca de cero está la media de las diferencias, más parecidas son las drogas en los efectos que se evalúan. El valor de t, que relaciona el promedio de las diferencias con su desviación estándar, es el criterio adecuado para someter a prueba la igualdad de los promedios (diferencia = 0).

Para resolver esto, se recurre al procedimiento MEANS. El siguiente programa realiza la comparación de las dos drogas:

```
DATA Drogas;
INPUT DrogaA DrogaB @@;
Diferen = DrogaA - DrogaB;
CARDS;
2.0 3.0 1 2 0 1
2.0 3.0 1 3 2 4
1.3 3.8 2 2 3 4
2 4 1.0 3 0 3
;
PROC MEANS MEAN T PRT;
   VAR Diferen;
RUN;
```

En el programa se creó la variable Diferen, para calcular la diferencia entre las dos drogas. Las opciones usadas en el PROC MEANS son: MEAN (media), T (el valor de "t") y PRT (la probabilidad de que la media sea cero).

La salida del programa anterior es la siguiente:

```
                    The MEANS Procedure

                Analysis Variable : Diferen

        Mean        t Value       Pr > |t|
-------------------------------------------
 -1.5416667          -6.37          <.0001
-------------------------------------------
```

La salida muestra la media de las diferencias, el valor de t calculado y el nivel de significancia. La probabilidad de obtener un valor mayor que T = -6.37 es muy baja (0.0001), lo cual indica que la diferencia es altamente significativa. Se puede concluir que la efectividad de las drogas es diferente con 99.9% de seguridad. Por lo tanto, se rechaza H_0: $\mu=0$.

Práctica

Se evaluaron dos dietas para alimentación de cerdos. Las ganancias de peso (kg) para los animales observados fueron los siguientes:

DIETA 1	DIETA 2
1.2	0.8
0.8	0.4
1.4	0.3
1.7	0.6
2.3	0.9

Con los datos anteriores, realice una prueba para determinar si la ganancia de peso de los animales son iguales para ambas dietas. Utilice un alpha del 5%

Pruebas No Paramétricas

Si los supuestos acerca de la distribución de los datos, por ejemplo que estos tengan una distribución normal, no se cumplen, o se tienen serias dudas acerca de su distribución, las pruebas estadísticas que se basan en estos supuestos, denominadas pruebas paramétricas podrían no ser válidas. Una alternativa es la de aplicar pruebas no paramétricas, las cuales no dependen de la forma o distribución de los datos.

Existen muchas pruebas no paramétricas que se aplican a diferentes situaciones. El SAS cuenta con varios procedimientos que hacen uso de métodos no paramétricos, entre estos tenemos los procedimientos: NPAR1WAY, FREQ, UNIVARIATE, CORR, KDE. En este documento se estudiará el PROC NPAR1WAY ya que es el procedimiento que hace las pruebas no paramétricas mas utilizadas.

El PROC NPAR1WAY realiza pruebas no paramétricas para localización y diferencias de escala para clasificaciones a una vía (una sola fuente de variación), para dos o mas muestras independientes.

La forma básica del PROC NPAR1WAY es:

```
PROC NPAR1WAY opciones;
     CLASS variable clasificatoria;
     VAR variables de respuesta;
```

Entre las opciones del NPAR1WAY se destacan:

- **ANOVA**: hace el análisis de variancia estándar a los datos.

- **WILCOXON**: cuando la variable clasificatoria tiene solo dos niveles hace la Prueba de Wilcoxon (Mann – Whitney). Para cualquier numero de niveles de la variable clasificatoria, siempre hace la prueba de Kruskall - Wallis.

Ejemplo: En un experimento se prueban dos variedades de chile y en un conteo de número de insectos por planta, realizado en quince plantas de cada variedad se obtiene:

Var. 1 :	5	7	8	12	2	8	23	6	8	9	2	5	14	7	21
Var. 2 :	21	8	2	23	7	9	3	24	7	2	1	12	7	8	9

El programa SAS es el siguiente:

```
DATA CHILE;
INPUT variedad conteo @@;
CARDS;
1  5  2 21 1  7  2  8
1  8  2  2 1 12  2 23
1  2  2  7 1  8  2  9
1 23  2  3 1  6  2 24
1  8  2  7 1  9  2  2
1  2  2  1 1  5  2 12
1 14  2  7 1  7  2  8
1 21  2  9
;
PROC NPAR1WAY WILCOXON;
CLASS variedad;
VAR conteo;
RUN;
```

La salida se muestra a continuacion:

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable conteo
Classified by Variable variedad

variedad	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	15	230.50	232.50	23.953079	15.366667
2	15	234.50	232.50	23.953079	15.633333

Average scores were used for ties.

Wilcoxon Two-Sample Test

Statistic	230.5000
Normal Approximation	
Z	-0.0626
One-Sided Pr < Z	0.4750
Two-Sided Pr > |Z|	0.9501
t Approximation	
One-Sided Pr < Z	0.4752
Two-Sided Pr > |Z|	0.9505

Z includes a continuity correction of 0.5.

Kruskal-Wallis Test

Chi-Square	0.0070
DF	1
Pr > Chi-Square	0.9335

Al detectar el SAS que la variable variedad, que se indica en la instrucción CLASS tiene dos niveles entonces se realizan tres pruebas: la Prueba de Wilcoxon (Mann – Whitney), la Prueba t aproximada (paramétrica) para comparar dos muestras independientes y la Prueba de Kruskall – Wallis. Tanto la Prueba de Wilcoxon como la Prueba t se hacen a una y a dos colas. Además, la Prueba de Wilcoxon por defecto tiene una corrección por continuidad de 0.5. De la Prueba de Wilcoxon, tanto a una como a dos colas se concluye que las dos variedades no difieren en ubicación con respecto al numero de insectos por planta (Pr<z=0.4750, Pr>|z|=0.9501). De la Prueba t se concluye lo mismo (Pr<z=0.4752, Pr>|z|=0.9505). La Prueba de Kruskall – Wallis también concluye que las dos variedades no difieren en ubicación con respecto al número de insectos por planta (Pr>Chi-cuadrado=0.9335).

Ejemplo: Catorce pollos fueron divididos en tres grupos de tamaños n1=n2=5, n3=6 y asignados a tres dietas diferentes. Los pesos en gramos después de un mes fueron:

Dieta	Pesos en gramos					
1	123	121	159	138	178	
2	144	172	165	143	179	
3	139	146	161	149	140	126

El programa SAS seria el siguiente:

```
DATA A;
INPUT dieta peso @@;
CARDS;
1 123 2 144 3 139
1 121 2 172 3 146
1 159 2 165 3 161
1 138 2 143 3 149
1 178 2 179 3 140
            3 126
;
PROC NPAR1WAY WILCOXON ANOVA;
CLASS dieta;
VAR peso;
RUN;
```

La salida del programa se muestra a continuación:

The NPAR1WAY Procedure

Analysis of Variance for Variable peso
Classified by Variable dieta

dieta	N	Mean
1	5	143.80
2	5	160.60
3	6	143.50

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Among	2	989.43750	494.718750	1.5529	0.2485
Within	13	4141.50000	318.576923		

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable peso
Classified by Variable dieta

dieta	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	5	33.0	42.50	8.827042	6.60
2	5	58.0	42.50	8.827042	11.60
3	6	45.0	51.00	9.219544	7.50

Kruskal-Wallis Test

Chi-Square	3.1809
DF	2
Pr > Chi-Square	0.2038

El SAS detecta que la variable dieta en la instrucción CLASS tiene tres niveles por lo cual solo hace la Prueba de Kruskall – Wallis. Esta prueba indica que no existe diferencia en la ubicación de las tres dietas, o sea que el peso de los pollos sometidos a las tres dietas es igual (Pr>Chi-cuadrado=0.2038). Al usar la instrucción ANOVA, se muestra al inicio de la salida el análisis de variancia, con cual se llega a la misma conclusión (Pr>F=0.2485).

Práctica

Los siguientes datos, representan el conteo del número de insectos por planta de una plaga en particular que ataca al cultivo de tomate, para tres diferentes variedades de tomate:

Variedad	Planta										
	1	2	3	4	5	6	7	8	9	10	11
1	1	12	3	4	2	11	12	3	4	11	5
2	12	9	10	11	14	15	13	12	13		
3	2	4	7	5	4	3	12	4	12	7	

Resolver los siguientes aspectos:

a. **Hacer un programa SAS que lea los datos.**

b. **Haga la prueba de Kuskall-Wallis para probar si existen diferencias en cuanto al número de insectos por planta en las tres variedades.**

c. **En el mismo programa, haga la prueba de Wilcoxon (Mann-Whitney) para comparar todas las variedades entre si por pares.**

d. **Interprete los resultados, para esto use un nivel de significancia del 5% en todas las pruebas.**

Comparación de Varias Medias: Análisis de Varianza

Si se desea hacer inferencia estadística de más de dos medias, se debe utilizar el análisis de varianza. El sistema SAS ofrece varios procedimientos para hacer análisis de varianza: **PROC GLM, PROC ANOVA** y **PROC NESTED**. En este documento se estudiará el PROC GLM, ya que es el procedimiento en SAS más utilizado para análisis de varianza debido a que:

- Permite trabajar con diseños no balanceados y balanceados
- Puede realizar análisis de covarianza
- Se pueden calcular contrastes ortogonales, ya sea para comparar grupos de tratamientos o para determinar curvas de respuesta.

A continuación la forma básica del PROC GLM:

```
PROC GLM;
     CLASSES Fuentes de variación;
     MODEL Variables de respuesta = Variables independientes;
```

Dependiendo del diseño y de las comparaciones se pueden incluir las siguientes instrucciones;

TEST H=Efectos **E**=Efecto; (Cuando se tiene más de un error)

MEANS Efectos / Comparación; (para comparaciones múltiples de medias: DUNCAN, TUKEY, LSD, BONFERRONI, DUNNET, SCHEFFE)

LSMEANS Efectos / **PDIFF**; Calcula medias ajustadas a los efectos. También calcula pruebas de T individuales para interacciones presentando una matriz de todas las posibles combinaciones.

El análisis de varianza, parte la variación total dentro de las observaciones en porciones asociadas con ciertos factores, los cuales están definidos por el esquema de clasificación de los datos. Estos factores son las fuentes de variación. Por ejemplo, la variación en la producción de leche en vacas puede ser fraccionada en porciones asociadas con diferencias entre hatos, diferencias genéticas y otras diferencias. Estas divisiones se hacen en términos de sumas de cuadrados asociadas a los grados de libertad. A continuación se presentan los diseños experimentales más utilizados en el campo agropecuario y de recursos naturales y la forma en que se analizan en SAS.

Diseño Completamente al Azar

Este tipo de diseño experimental, asume que las unidades experimentales de una población se asignan al azar a grupos que generalmente son llamados tratamientos. La hipótesis nula es que las poblaciones estudiadas tienen la misma media. Por ejemplo, se tienen datos de cuatro variedades de arroz y se quiere saber si se debe aceptar que los promedios de ellas (en la población) no difieren entre sí. Los rendimientos (en kgs) observados de cuatro parcelas de cada variedad son:

Variedad	**Rendimiento**			
1	23.5	31.2	18.2	27.8
2	24.4	26.3	28.2	28.3
3	32.3	28.3	33.0	34.5
4	38.9	42.1	32.3	38.2

El modelo matemático ajustado a estos datos sería:

$$\textbf{Rendimiento} = \mu + V_i + e_{ij}$$

Donde:

μ es el promedio de las medias de las variedades

V_i es el efecto de la Variedad $_i$

e_{ij} es el error experimental

Este análisis de varianza puede hacerse, lo mismo que comparaciones múltiples de las medias de las variedades, usando PROC GLM con la opción para la prueba de DUNCAN. Esto se haría con el siguiente programa:

```
DATA Varianza;
DO Variedad=1 TO 4
        INPUT Rend @@;
        OUTPUT;
END;
CARDS;
23.5 24.4 32.3 38.9 31.2 26.3 28.3 42.1
18.2 28.2 33.0 32.3 27.8 28.3 34.5 38.2
;
PROC GLM;
      CLASS Variedad;
      MODEL Rend = Variedad;
      MEANS Variedad / DUNCAN;
RUN;
```

Los datos están clasificados sólo de acuerdo a los valores de variedad; por lo tanto, variedad es la única variable que debe aparecer en la instrucción CLASS. La variable Rendimiento es la variable de respuesta que va a analizarse; por lo tanto, Rend aparece al lado izquierdo del signo '=' en la línea MODEL. La única fuente de variación además del ERROR (residuo), es Variedad; por lo tanto, se digita variedad a la derecha del signo "=". La salida producida por la instrucción MODEL es la siguiente:

```
                          The GLM Procedure

                     Class Level Information
                Class          Levels    Values
                Variedad            4    1 2 3 4

                  Number of observations    16

Dependent Variable: Rend

                                   Sum of
Source                  DF         Squares      Mean Square    F Value    Pr > F

Model                    3     395.0318750 (1)   31.6772917       8.98    0.0022
Error                   12     175.9825000       14.6652083
Corrected Total         15     571.0143750

              R-Square     Coeff Var      Root MSE     Rend Mean
              0.691807 (2)  12.56867 (3)  3.829518      30.46875

Source                  DF       Type I SS     Mean Square    F Value    Pr > F

Variedad                 3     395.0318750 (4) 131.6772917       8.98    0.0022 (5)

Source                  DF     Type III SS     Mean Square    F Value    Pr > F

Variedad                 3     395.0318750     131.6772917       8.98    0.0022
```

Note que la suma de cuadrados del Modelo **(1)**, es la misma que la suma de cuadrados de Variedad **(4)**. Esto se debe a que la variedad es la única fuente de variación en el modelo, aparte del ERROR. Recuerde que la suma de las sumas de cuadrados de las fuentes de variación, es igual a la suma de cuadrados total. El valor de Pr > P de 0.0022 **(5)** indica que hay diferencias altamente significativas entre las medias de las variedades. El valor de r-cuadrado es de 0.691807 **(2)**, lo que indica que el 69.18% de los datos se ajustan al modelo. El coeficiente de variación es 12.56 **(3)**. Recuerde que coeficientes de variación mayores que 20, indican que existen problemas con el modelo que se está aplicando a los datos.

Los resultados del PROC GLM se pueden resumir en la siguiente tabla:

Fuente de Variación	Grados de Libertad	Suma de Cuadrados	Cuadrado Medio	F Calculada	Pr > F
Variedades	3	395.0318	31.6773	8.98	0.0022
Error	12	175.9825	14.6652		
Total	15	571.0143			

La instrucción MEANS (para realizar prueba de Duncan a las medias) produce la siguiente salida:

```
                Duncan's Multiple Range Test for Rend

NOTE: This test controls the Type I comparisonwise error rate, not the
experimentwise error rate.

                    Alpha                          0.05
                    Error Degrees of Freedom         12
                    Error Mean Square           14.66521


          Number of Means          2          3          4
          Critical Range       5.900      6.176      6.343


      Means with the same letter are not significantly different.

       Duncan Grouping          Mean      N    Variedad

                    A        37.875      4    4
                    A
               B    A        32.025      4    3
               B
               B    C        26.800      4    2
                    C
                    C        25.175      4    1
```

N indica el número de observaciones usadas para calcular cada media. Como lo dice la salida de SAS, las medias que tengan la misma letra en la columna GROUPING no son significativamente diferentes. En este caso las diferencias encontradas son: Variedad 4 es diferente a Variedades 2 y 1, Variedad 3 es diferente a Variedad 1.

Algo importante que se debe tener en cuenta, es que la mayoría de las pruebas de comparación múltiple requieren que los tratamientos tengan el mismo núnero de observaciones.

Con frecuencia, es más adecuado utilizar contrastes en vez de pruebas de comparaciones múltiples. Además, si los tratamientos son de tipo cuantitativo en lugar de cualitativo (ej. niveles de algún fertilizante), en vez de usar una prueba de comparacion múltiple, es preferible ajustar algún modelo de regresión a los datos (Ver "Análisis de varianza con tratamientos cuantitativos" página 117).

Diseño de Bloques Completos al Azar

Este diseño se emplea cuando se supone que las unidades experimentales pueden formar grupos relativamente homogéneos, generalmente llamados bloques. Los tratamientos se asignan al azar a las unidades experimentales dentro de los bloques.

Ejemplo: se realizó un experimento para evaluar cinco variedades de maíz, las cuales se repitieron en 4 bloques. Los datos, en kg/parcela obtenidos son los siguientes:

Bloque	**Tratamiento**				
	1	**2**	**3**	**4**	**5**
1	25.4	34.3	23.4	28.3	12.2
2	34.2	38.8	28.2	28.2	14.2
3	21.2	32.1	21.2	26.2	14.5
4	39.9	45.8	34.3	32.3	19.3

Se emplea el siguiente modelo para cualquier rendimiento Yij perteneciente al bloque i y al tratamiento j:

$$Y_{ij} = \mu + B_i + T_j + e_{ij}$$

Donde:

μ es la media general de los tratamientos

B_i es el efecto del bloque i

T_j es el efecto del tratamiento j

e_{ij} es el error experimental

La tabla del análisis de varianza para este experimento es:

Fuentes de variación	**GL**
Bloques	3
Tratamientos	4
Error	12
Total	19

El programa en SAS para analizar los datos se presenta a continuación:

```
DATA Varianza;
DO Bloque = 1 TO 4;
     DO Trat = 1 TO 5;
              INPUT  Rend @@;
              OUTPUT;
      END;
END;
CARDS;
25.4   34.3   23.4   28.3   12.2
34.2   38.8   28.2   28.2   14.2
21.2   32.1   21.2   26.2   14.5
39.9   45.8   34.3   32.3   19.3
;
PROC GLM;
     CLASS Bloque Trat;
     MODEL Rend = Bloque Trat;
     MEANS Trat / DUNCAN;
RUN;
```

En la instrucción CLASS, se especifican las variables que identifican los criterios de clasificación que corresponden al diseño. En este caso, los datos están clasificados de acuerdo al bloque y al tratamiento al que pertenecen. La variable de respuesta es el rendimiento (Rend), la cual debe ir a la izquierda del signo "=". En este ejemplo, las dos fuentes de variación además del Error son Bloque y Trat, por lo cual deben aparecer a la derecha del signo "=".

Los resultados del análisis producidos por el programa son los siguientes:

The GLM Procedure

Class Level Information

Class	Levels	Values
Bloque	4	1 2 3 4
Trat	5	1 2 3 4 5

Number of observations 20

```
Error                  12        94.501000         7.875083
Corrected Total        19      1548.240000

                R-Square      Coeff Var      Root MSE      Rend Mean

                0.938962       10.13089      2.806258       27.70000

Source                 DF       Type I SS     Mean Square    F Value    Pr > F

Bloque                  3      377.304000      125.768000      15.97    0.0002
Trat                    4     1076.435000      269.108750      34.17    <.0001

Source                 DF      Type III SS     Mean Square    F Value    Pr > F

Bloque                  3      377.304000      125.768000      15.97    0.0002
Trat                    4     1076.435000      269.108750      34.17    <.0001
```

Los resultados del PROC GLM se pueden resumir en la siguiente tabla:

Fuente de Variación	Grados de Libertad	Suma de Cuadrados	Cuadrado Medio	F Calculada	Pr > F
Bloques	3	377.3040	125.768	15.97	0.0002
Tratamientos	4	1076.435	269.109	34.17	0.0001
Error	12	94.501	7.875		
Total	19	1548.240			

De estos resultados se concluye que los bloques son significativamente diferentes; esto indica que la formación de bloques surtió efecto. También se puede concluir que hay diferencias reales entre tratamientos.

El resultado de la instrucción MEANS (compración de medias) es el siguiente:

```
                        The GLM Procedure

                Duncan's Multiple Range Test for Rend

NOTE: This test controls the Type I comparisonwise error rate, not the
experimentwise error rate.

                Alpha                          0.05
                Error Degrees of Freedom         12
                Error Mean Square          7.875083


      Number of Means          2          3          4          5
      Critical Range       4.323      4.525      4.648      4.729


      Means with the same letter are not significantly different.

          Duncan Grouping          Mean      N     Trat

                       A        37.750      4     2

                       B        30.175      4     1
                       B
                       B        28.750      4     4
                       B
                       B        26.775      4     3

                       C        15.050      4     5
```

En este caso, la prueba de Duncan indica que la media del tratamiento 2 es diferente a las demás. En el extremo inferior, la media del tratamiento 5 también es diferente del resto. Las medias de los tratamientos 1, 3 y 4 no son diferentes entre sí, pero sí diferentes a la media del tratamiento 2 y del 5.

Si el experimento consta de tratamientos cuantitativos (Ej. niveles de algún fertilizante.)

Ver "Análisis de varianza con tratamientos cuantitativos" página 117.

Diseño de Cuadrado Latino

En el diseño de bloques al azar, se busca aislar una fuente de variación extraña reconocible usando bloques. Si el bloqueo funciona, el cuadrado medio del error se reduce, dando una prueba más sensible que la obtenida con el diseño completamente al azar.

No obstante, algunas veces hay gradientes en dos sentidos que deben ser aislados formando bloques en dos direcciones. Por ejemplo, en un campo, los gradientes de fertilidad pueden existir en dos sentidos: paralelamente y en ángulo recto a los surcos del arado. El empleo del diseño de bloques al azar aísla solamente una de estas fuentes de variación, mientras la otra forma parte del término de error, lo cual reduce la precisión de la prueba.

Supóngase que se tienen cinco tratamientos (A,B,C,D,E) en un diseño de cuadrado latino, distribuídos de la siguiente manera:

	COLUMNAS				
FILAS	1	2	3	4	5
1	C (32)	A (23)	B (24)	E (34)	D (28)
2	A (28)	C (36)	D (33)	B (32)	E (38)
3	E (40)	D (37)	A (31)	C (39)	B (26)
4	D (39)	B(31)	E (43)	A (33)	C (41)
5	B (32)	E (45)	C (43)	D (40)	A (35)

Los valores a la derecha de cada letra (tratamientos) son los rendimientos de maíz por parcela.

El siguiente programa realiza el análisis de varianza en cuadrado latino a los datos anteriores:

```
DATA Latino;
DO Fila = 1 TO 5;
    DO Columna = 1 TO 5;
            INPUT Trat $ Rend  @@;
            OUTPUT;
     END;
END;
CARDS;
C 32 A 23 B 24 E 34 D 28 A 28 C 36 D 33 B 32 E 38
E 40 D 37 A 31 C 39 B 26 D 39 B 31 E 43 A 33 C 41
B 32 E 45 C 43 D 40 A 35
;
PROC GLM;
    CLASS Columna Fila Trat;
    MODEL Rend = Columna Fila Trat;
    MEANS Trat / DUNCAN;
RUN;
```

La salida del programa anterior se muestra a continuación:

```
                              The GLM Procedure

                           Class Level Information

                  Class          Levels    Values
                  Columna             5    1 2 3 4 5
                  Fila                5    1 2 3 4 5
                  Trat                5    A B C D E

Dependent Variable: Rend

                                                Sum of
Source                       DF           Squares      Mean Square    F Value    Pr > F
Model                        12      835.1200000       69.5933333      26.84    <.0001
Error                        12       31.1200000        2.5933333
Corrected Total              24      866.2400000

                       R-Square     Coeff Var      Root MSE      Rend Mean
                       0.964075      4.665072      1.610383       34.52000

Source                       DF        Type I SS     Mean Square    F Value    Pr > F

Columna                      4       11.0400000       2.7600000       1.06    0.4159
Fila                         4      347.8400000      86.9600000      33.53    <.0001
Trat                         4      476.2400000     119.0600000      45.91    <.0001

Source                       DF       Type III SS     Mean Square    F Value    Pr > F

Columna                      4       11.0400000       2.7600000       1.06    0.4159
Fila                         4      347.8400000      86.9600000      33.53    <.0001
Trat                         4      476.2400000     119.0600000      45.91    <.0001
```

Los resultados del PROC GLM se pueden resumir en la siguiente tabla:

Fuente de Variación	Grados de Libertad	Suma de Cuadrados	Cuadrado Medio	F Calculada	Pr > F
Columna	4	11.04	2.76	1.06	0.4159
Fila	4	347.84	86.96	33.53	0.0001
Tratamiento	4	476.24	119.06	45.91	0.0001
Error	12	31.12	2.59		
Total	24	866.24			

Como se puede ver, el efecto de fila es altamente significativo, lo mismo que los tratamientos. En este caso, no se ganó precisión al emplear columnas como fuente de variación.

La salida de la prueba de Duncan para los tratamientos es la siguiente:

Duncan's Multiple Range Test for Rend

NOTE: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	12
Error Mean Square	2.593333

Number of Means	2	3	4	5
Critical Range	2.219	2.323	2.386	2.427

Means with the same letter are not significantly different.

Duncan Grouping	Mean	N	Trat
A	40.000	5	E
A	38.200	5	C
B	35.400	5	D
C	30.000	5	A
C			
C	29.000	5	B

De la prueba de Duncan, se ve que las medias de los tratamientos E y C no difieren entre sí, pero son diferentes a las medias D, A y B. La media D, es diferente a todas las demás. Las medias de A y B son iguales entre sí, pero diferentes a las medias de E, C y D.

Práctica

Con el propósito de evaluar el comportamiento de cinco variedades de frijol (A, B, C, D y E), se instaló un experimento bajo condiciones de invernadero utilizando un diseño completamente al azar en el cual los tratamientos se repitieron 6 veces. La variable respuesta fue el peso de materia seca (g) 30 días después de la siembra. Los resultados se presentan en el cuadro siguiente:

VARIEDADES Y PESOS SECOS				
A 2.9	B 3.3	C 3.1	D 4.5	E 6.6
A 3.5	B 3.8	C 3.8	D 4.4	E 6.8
A 4.1	B 3.7	C 4.2	D 4.8	E 7.4
A 3.0	B 3.6	C 3.1	D 4.7	E 7.2
A 3.0	B 3.1	C 3.5	D 4.5	E 7.0
A 3.6	B 3.3	C 3.2	D 5.0	E 7.6

a. **Realizar un análisis de varianza**
b. **Realizar las pruebas de Duncan a las medias de peso seco de las variedades a niveles de significancia de 1% y 5%**
c. **Concluir con base a los resultados obtenidos**

Se compraró el efecto de varios herbicidas sobre el peso de las flores de gladiolos. El peso promedio de inflorescencia en onzas se da a continuación para los cuatro tratamientos. El deseño utilizado es en bloques completos al azar.

TRATAMIEN-	BLOQUES			
TOS	1	2	3	4
Control	1.25	1.73	1.82	1.31
2.4-D TCA	2.05	1.56	1.68	1.69
DN/Cr	1.95	2.00	1.83	1.81
Sesin	1.75	1.93	1.70	1.59

Realizar un análisis de varianza a los datos anteriores, de acuerdo al diseño utilizado. Aplique una prueba de Duncan, Tukey y Dunnet a las medias de los tratamientos y compare las tres pruebas.

Concluir con base a los resultados obtenidos

Arreglos Factoriales

Los arreglos factoriales permiten estudiar varios factores simultáneamente con muy poco trabajo adicional; aumentan la precisión, la cobertura y utilidad de los resultados al proveer información sobre las interacciones entre los factores en prueba.

Como ilustración, considérese un arreglo factorial que envuelve tres variedades de caña de azúcar (V) y tres niveles de nitrógeno (N), conducido usando un diseño de bloques completos al azar con dos repeticiones.

Cuando se analizan los resultados del experimento se pueden hacer las siguientes comparaciones:

i) comparaciones entre variedades
ii) comparaciones entre niveles de nitrógeno
iii) la interacción entre variedades y nitrógeno

Las comparaciones i y ii son entre efectos principales. La presencia o ausencia de efectos principales no dice nada acerca de la presencia o ausencia de la interacción o viceversa; por lo tanto, se deben considerar separadamente.

Una interacción significativa implica que los efectos de los factores no son independientes entre sí. En este caso, no se puede concluir que el mejor tratamiento corresponde a la combinación de la variedad con el mayor promedio y el nivel de nitrógeno con el promedio más alto. Es necesario estudiar más a fondo cómo se comporta cada variedad con los diferentes niveles de fertilización, o los niveles de fertilización con cada variedad. A continuación los datos de rendimiento en toneladas por hectárea:

		VARIEDADES		
	FERTILIZACION	V0	V1	V2
Bloque I	F0	66.52	61.45	68.60
	F1	68.98	62.55	64.54
	F2	75.95	57.90	68.09
Bloque II	F0	56.50	53.45	58.50
	F1	58.95	51.55	54.54
	F2	66.95	47.90	58.19

El modelo matemático para este factorial es el siguiente:

$$Y_{ijk} = \mu + B_i + V_j + F_k + VF_{jk} + e_{ijk}$$

Donde:

μ	es la media general
B_i	es el efecto del bloque i
V_j	es el efecto principal de la variedad j
F_k	es el efecto principal del nivel k de fertiliz.
VF_{jk}	es la interacción de la variedad j por el nivel k en el bloque i
e_{ijk}	es el error experimental

La tabla del análisis de varianza en este caso sería:

Fuentes de Variación	**GL**
Bloque	1
Variedad	2
Fertilizante	2
Variedad*Fertilizante	4
Error	8
Total	17

Puede emplearse al siguiente programa SAS para analizar los datos:

```
DATA Factoria;
DO Bloque = 1 TO 2;
   DO Fertiliz = 'F0', 'F1', 'F2';
      DO Variedad = 'V0','V1','V2';
         INPUT Rend @@;
         OUTPUT;
      END;
   END;
END;
CARDS;
66.52 61.45  68.60 68.98  62.55  64.54 75.95  57.90  68.09 56.50  53.45
58.50 58.95  51.55  54.54 66.95  47.90  58.19
;
PROC GLM;
CLASS Bloque Variedad Fertiliz;
MODEL Rend = Bloque Variedad Fertiliz Variedad*Fertiliz;
LSMEANS Variedad*Fertiliz / PDIFF;
RUN;
```

*En este programa se incluye la instrucción **LSMEANS** con la opción **PDIFF** para realizar la prueba de T a la interacción Variedades por Fertilizante.*

Los resultados del programa anterior son los siguientes:

The GLM Procedure

Class Level Information

Class	Levels	Values
Bloque	2	1 2
Variedad	3	V0 V1 V2
Fertiliz	3	F0 F1 F2

Number of observations 18

The GLM Procedure

Dependent Variable: Rend

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	880.7389611	97.8598846	279.66	<.0001
Error	8	2.7994000	0.3499250		
Corrected Total	17	883.5383611			

R-Square	Coeff Var	Root MSE	Rend Mean
0.996832	0.967006	0.591545	61.17278

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Bloque	1	430.7112500	430.7112500	1230.87	<.0001
Variedad	2	297.9283444	148.9641722	425.70	<.0001
Fertiliz	2	17.0481444	8.5240722	24.36	0.0004
Variedad*Fertiliz	4	135.0512222	33.7628056	96.49	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Bloque	1	430.7112500	430.7112500	1230.87	<.0001
Variedad	2	297.9283444	148.9641722	425.70	<.0001
Fertiliz	2	17.0481444	8.5240722	24.36	0.0004
Variedad*Fertiliz	4	135.0512222	33.7628056	96.49	<.0001

Aunque los efectos principales, Variedad (0.0001) y Fertilizante (0.0004) son significativos, no se pueden estudiar por separado, ya que la mejor dosis de fertilizante depende de la variedad que se este investigando, según debe inferirse del hecho de que la interacción también es significativa (0.0001). Lo más recomendable en este caso, sería hacer un análisis estudiando cuál es la tendencia de la respuesta de cada variedad a la fertilización. Para ello, se calculan los componentes lineal y cuadrático para cada variedad, empleando dos grados de libertad por variedad. Los seis grados de libertad que se ocupan corresponden a los de fertilizantes, más los de la interacción; en forma correspondiente, la suma de las sumas de cuadrados de los componentes es igual a la suma de las sumas de cuadrados de Fertilizante y Variedad*Fertilizante.

Si la interacción no hubiera sido significativa, se hubiera podido hacer comparaciones de las medias de las variedades y ajustar curvas de respuesta para los niveles de fertilización del promedio de las variedades. Las pruebas de comparación múltiple (LSMEANS) para la intereacción calculadas por el programa se presentan a continuación:

The GLM Procedure
Least Squares Means

Variedad	Fertiliz	Rend LSMEAN	LSMEAN Number
V0	F0	61.5100000	1
V0	F1	63.9650000	2
V0	F2	71.4500000	3
V1	F0	57.4500000	4
V1	F1	57.0500000	5
V1	F2	52.9000000	6
V2	F0	63.5500000	7
V2	F1	59.5400000	8
V2	F2	63.1400000	9

Least Squares Means for effect Variedad*Fertiliz
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: Rend

i/j	1	2	3	4	5	6	7	8	9
1		0.0032	<.0001	0.0001	<.0001	<.0001	0.0087	0.0104	0.0248
2	0.0032		<.0001	<.0001	<.0001	<.0001	0.5029	<.0001	0.2006
3	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
4	0.0001	<.0001	<.0001		0.5180	<.0001	<.0001	0.0077	<.0001
5	<.0001	<.0001	<.0001	0.5180		0.0001	<.0001	0.0030	<.0001
6	<.0001	<.0001	<.0001	<.0001	0.0001		<.0001	<.0001	<.0001
7	0.0087	0.5029	<.0001	<.0001	<.0001	<.0001		0.0001	0.5079
8	0.0104	<.0001	<.0001	0.0077	0.0030	<.0001	0.0001		0.0003
9	0.0248	0.2006	<.0001	<.0001	<.0001	<.0001	0.5079	0.0003	

NOTE: To ensure overall protection level, only probabilities associated with pre-plannedcomparisons should be used.

Como muestra la salida del LSMEANS, se obtienen 9 medias ajustadas, ya que cada media representa la interacción Variedad * Fertilizante (3 x 3). Se muestra una matriz de 9 X 9, que muestra la probabilidad asociada a T de que las medias sean iguales. En este caso, cuando el valor de la probabilidad es igual a <.0001 , se rechaza H_0: *las medias son iguales*. Esto no implica que se pueda trabajar con otro alfa, por ejemplo, 0.05. Esta decisión es responsabilidad del investigador.

Diseño de parcelas divididas

La necesidad de utilizar el diseño de parcelas divididas, surge cuando se aplican dos o más tipos de tratamientos en arreglos factoriales, si los niveles de un factor puede aplicarse a parcelas relativamente pequeñas, mientras que los del otro, sea más conveniente aplicarlos a parcelas más grandes.

Un ejemplo del uso del diseño de parcelas divididas es cuando se prueban diferentes niveles de irrigación en las parcelas grandes y factores tales como variedades o fertilizantes son aplicados a las parcelas pequeñas. Supóngase que se tiene un experimento con dos niveles de irrigación (alta y moderada) y cuatro variedades de caña en cuatro bloques. Los datos de los rendimientos de la caña son:

		Variedad			
	Irrigación	**1**	**2**	**3**	**4**
Bloque I	alta	123.2	132.3	123.2	128.8
	moderada	118.2	123.2	115.2	116.3
Bloque II	alta	128.2	138.3	128.2	125.8
	moderada	119.2	120.2	117.2	121.3
Bloque III	alta	118.2	122.3	121.2	124.8
	moderada	111.2	117.2	113.2	113.3
Bloque IV	alta	128.2	123.3	128.2	132.8
	moderada	113.2	122.2	114.2	116.3

El modelo matemático para este experimento es:

$$Y_{ijk} = \mu + B_i + I_j + e_{ij} + V_i + IV_{ij} + e_{ijk}$$

Donde:

μ	es la media general
B_i	es el efecto del bloque i
I_j	es el efecto del factor de la parcela principal (Irrigación)
e_{ij}	es el error experimental asociado a la parcela grande (Irrigación)
V_i	es el efecto del factor de la subparcela (Variedad)
IV_{ij}	es la interacción entre método de irrigación$_j$ y variedad$_k$
e_{ijk}	es el error experimental asociado a las subparcelas

La tabla del análisis de varianza para este experimento sería:

Fuentes de variación	**GL**
Bloque	3
Irrigación	1
Error A (bloque*Irr)	3
Variedad	3
Variedad*Irrigación	3
Error B (Bloque*Irr + Var*Iirr*Bloque)	18
Total	31

El Error A es usado para probar las diferencias entre bloques e irrigaciones, y el error B es usado para probar las diferencias entre variedades y la interacción entre variedades e irrigaciones.

El programa en SAS sería el siguiente:

```
DATA Dividida;
DO Bloque = 1 TO 4;
   DO Irriga = 'Alta     ', 'Moderada';
      DO Variedad = 1 TO 4;
         INPUT Rend @@;
         OUTPUT;
      END;
   END;
END;
CARDS;
123.2          132.3          123.2          128.8
118.2          123.2          115.2          116.3
128.2          138.3          128.2          125.8
119.2          120.2          117.2          121.3
118.2          122.3          121.2          124.8
111.2          117.2          113.2          113.3
128.2          123.3          128.2          132.8
113.2          122.2          114.2          116.3
;
PROC GLM;
CLASSES Bloque Irriga Variedad;
MODEL Rend = Bloque Irriga Bloque*Irriga Variedad Variedad*Irriga;
TEST H=Bloque Irriga E=Bloque*Irriga;
RUN;
```

La instrucción TEST (última línea) indica que Bloque e Irriga deben probarse usando Bloque*Irriga como término de error. Si no se especifica esto, el programa ejecuta la prueba de F para Bloques y para Irrigación usando el error B , lo cual es incorrecto.

La salida de este programa sería la siguiente:

```
                    The GLM Procedure

                Class Level Information

Class          Levels    Values

Bloque              4    1 2 3 4
Irriga              2    Alta Moderada
Variedad            4    1 2 3 4


Number of observations    32
```

The GLM Procedure

Dependent Variable: Rend

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	1127.306250	86.715865	8.00	<.0001
Error	18	195.062500	10.836806		
Corrected Total	31	1322.368750			

R-Square	Coeff Var	Root MSE	Rend Mean
0.852490	2.702041	3.291930	121.8313

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Bloque	3	214.5937500	71.5312500	6.60	0.0033
Irriga	1	754.6612500	754.6612500	69.64	<.0001
Bloque*Irriga	3	18.0937500	6.0312500	0.56	0.6504
Variedad	3	129.9237500	43.3079167	4.00	0.0241
Irriga*Variedad	3	10.0337500	3.3445833	0.31	0.8188

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Bloque	3	214.5937500	71.5312500	6.60	0.0033
Irriga	1	754.6612500	754.6612500	69.64	<.0001
Bloque*Irriga	3	18.0937500	6.0312500	0.56	0.6504
Variedad	3	129.9237500	43.3079167	4.00	0.0241
Irriga*Variedad	3	10.0337500	3.3445833	0.31	0.8188

Tests of Hypotheses Using the Type III MS for Bloque*Irriga as an Error Term

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Bloque	3	214.5937500	71.5312500	11.86	0.0359
Irriga	1	754.6612500	754.6612500	125.13	0.0015

La salida de SAS puede resumirse en la siguiente tabla:

Fuente de Variación	Grados de Libertad	Suma de Cuadrados	Cuadrado Medio	F Calculada	Pr > F
Bloques	3	214.59	71.53	11.86	0.0359
Irriga	1	754.66	754.66	125.13	0.0015
Error A (Bloque* Irriga)	3	18.09	6.03	0.56	0.6504
Variedad	3	129.92	43.31	4.0	0.0241
Variedad*Irriga	3	10.03	3.34	0.31	0.8188
Error B (Bloque*Irriga*Variedad)	18	195.06	10.84		

Como se puede ver en esta tabla de ANDEVA, el error que se usa para la prueba de F para probar Bloques e Irrigación es el error A (es decir la interacción Bloque*Irrigación). Note que SAS produce dos tablas de ANDEVA, una donde prueba todas las fuentes de variación contra el Cuadrado Medio del Error y otra donde prueba la parcela grande contra el error A. Por lo tanto, el usuario debe modificar la tabla del ANDEVA como se muestra en la última tabla.

En este ejemplo hubo diferencias significativas entre irrigaciones (0.0015) y entre variedades (0.0241), pero no hubo significancias en la interacción de irrigación * variedad (0.8188). Por lo tanto, se puede hacer una prueba de Duncan para las medias de variedades y para las medias de irrigaciones independientemente. Esto se hace agregando las siguientes líneas al programa SAS:

```
MEANS Irriga / DUNCAN E=Bloque*Irriga;
MEANS Variedad / DUNCAN;
```

Práctica

Digite el programa utilizado para realizar el análisis de varianza en parcelas divididas. Agregue las instrucciones para realizar la prueba de Duncan a los efectos principales Irrigación y Variedad.

Determine donde se presentan las diferencias en cada efecto principal.

Análisis de varianza con tratamientos cuantitativos

Cuando los tratamientos usados en un experimento son cuantitativos, por ejemplo, diferentes niveles de un fertilizante aplicados a varias parcelas con algún cultivo, no tiene mucho sentido hacer comparaciones múltiples para las medias de los tratamientos.

En este caso, puede ser más útil averiguar cual es la tendencia de la respuesta y no simplemente si hay diferencias entre los diferentes niveles o cual de los niveles empleados es el que produce el mayor rendimiento. Para hacer esto, se ajustan curvas de respuesta para ver si la relación entre el nivel de fertilización y el rendimiento es lineal, cuadrático, cúbico, etc.

Ejemplo: Se tienen los siguientes datos de Rendimientos en Kg, obtenidos al aplicar cuatro niveles de nitrógeno (gramos por parcela) a varias parcelas de maíz, en cuatro bloques

Bloque	**Nivel de Nitrógeno**			
	0	**50**	**100**	**150**
1	25.4	34.3	36.4	28.3
2	34.2	42.8	43.2	36.2
3	21.2	32.1	34.2	23.2
4	39.9	45.8	43.3	41.3

Debido a que se tienen cuatro niveles de nitrógeno, sólo se pueden ajustar polinomios hasta tercer grado (efecto cúbico). Para hacer esto en SAS sería como sigue:

```
DATA Nitro;
INPUT Bloque Nit Rend @@;
CARDS;
1 000 25.4 1 050 34.3 1 100 36.4 1 150 28.3
2 000 34.2 2 050 42.8 2 100 43.2 2 150 36.2
3 000 21.2 3 050 32.1 3 100 34.2 3 150 23.2
4 000 39.9 4 050 45.8 4 100 43.3 4 150 41.3
;
PROC GLM;
  CLASS Bloque Nit;
  MODEL Rend = Bloque Nit;
  CONTRAST 'Lineal'     Nit -3  -1   1   3;
  CONTRAST 'Cuadratico' Nit  1  -1  -1   1;
  CONTRAST 'Cubico'     Nit -1   3  -3   1;
RUN;
```

En el caso de este experimento, los niveles de nitrógeno están igualmente espaciados. Para este caso, se puede recurrir a la siguiente tabla de polinomios ortogonales, la cual incluye los coeficientes hasta seis tratamientos. Esta tabla, da los coeficientes de acuerdo al número de tratamientos que se tengan; en este caso se utilizan los coeficientes para cuatro tratamientos.

No. Tratam.	Grados del Polinomio	Tratamientos total					
		T1	T2	T3	T4	T5	T6
2	1	-1	+1				
3	1	-1	0	+1			
	2	+1	-2	+1			
4	1	-3	-1	+1	+3		
	2	+1	-1	-1	+1		
	3	-1	+3	-3	+1		
5	1	-2	-1	0	+1	+2	
	2	+2	-1	-2	-1	+2	
	3	-1	+2	0	-2	+1	
	4	+1	-4	+6	-4	+1	
6	1	-5	-3	-1	+1	+3	+5
	2	+5	-1	-4	-4	-1	+5
	3	-5	+7	+4	-4	-7	+5
	4	+1	-3	+2	+2	-3	+1
	5	-1	+5	-10	+10	-5	+1

La salida del programa es la siguiente:

```
                    The GLM Procedure

                Class Level Information

Class          Levels    Values

Bloque              4    1 2 3 4
Nit                 4    0 50 100 150

Dependent Variable: Rend

                                        Sum of
Source                  DF         Squares     Mean Square    F Value    Pr > F

Model                    6     824.5450000     137.4241667      37.62    <.0001
Error                    9      32.8725000       3.6525000
Corrected Total         15     857.4175000

          R-Square     Coeff Var      Root MSE     Rend Mean

          0.961661      5.442938      1.911151      35.11250
```

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Bloque	3	572.0225000	190.6741667	52.20	<.0001
Nit	3	252.5225000	84.1741667	23.05	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Bloque	3	572.0225000	190.6741667	52.20	<.0001
Nit	3	252.5225000	84.1741667	23.05	0.0001

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Lineal	1	9.1125000	9.1125000	2.49	0.1487
Cuadratico	1	243.3600000	243.3600000	66.63	<.0001
Cubico	1	0.0500000	0.0500000	0.01	0.9094

Del resultado de las pruebas de los contrastes, se observa que el polinomio cuadrático **(0.0001)** es el que explica la relación entre rendimiento y nitrógeno. Los polinomios lineal **(0.1487)** y cúbico **(0.9094)** no son significativos. En el caso de que todos los polinomios fueran significativos se debe escoger el polinomio de mayor grado.

El siguiente paso a seguir sería ajustar una regresión cuadrática a las medias de cada nivel de nitrógeno. Esto se haría agregando las siguientes líneas al programa anterior:

```
PROC SORT; BY Nit;
PROC MEANS MEAN NOPRINT; BY Nit; VAR Rend;
OUTPUT OUT=Medias MEAN = Rend;
PROC GLM;
MODEL Rend = Nit Nit*Nit;
OUTPUT PREDICTED = Py;
PROC PLOT;
PLOT Rend*Nit= '*' Py*Nit='+' / OVERLAY VPOS=15 HPOS=30;
RUN;
```

Las dos últimas líneas hacen un gráfico de los datos observados superponiendo la línea de regresión del modelo ajustado (cuadrático).

La salida de estas instrucciones es la siguiente:

```
                              The GLM Procedure

                          Number of observations    4


Dependent Variable: Rend
                                                 Sum of
Source                     DF       Squares     Mean Square    F Value    Pr > F
Model                       2     63.11812500    31.55906250    2524.72    0.0141
Error                       1      0.01250000     0.01250000
Corrected Total             3     63.13062500

              R-Square     Coeff Var      Root MSE     Rend Mean
              0.999802      0.318415      0.111803      35.11250

Source                     DF      Type I SS     Mean Square    F Value    Pr > F
Nit                         1     2.27812500     2.27812500     182.25    0.0471
Nit*Nit                     1    60.84000000    60.84000000    4867.20    0.0091

Source                     DF     Type III SS     Mean Square    F Value    Pr > F
Nit                         1    62.50637755    62.50637755    5000.51    0.0090
Nit*Nit                     1    60.84000000    60.84000000    4867.20    0.0091

                                                    Standard
          Parameter          Estimate                Error     t Value    Pr > |t|

          Intercept      30.20000000        0.10897247      277.13      0.0023
          Nit             0.24750000        0.00350000       70.71      0.0090
          Nit*Nit        -0.00156000        0.00002236      -69.77      0.0091

                  Plot of Rend*Nit.  Symbol used is '*'.
                  Plot of Py*Nit.    Symbol used is '+'.
                     40 |
                        |                    *
                        |           *
                   Rend |
                        |
                        |
                        |
                     35 |
                        |
                        |
                        |                              *
                        |
                        |
                     30 |  *
                        --|--------|--------|--------|-
                          0        50       100      150
                                    Nit

          NOTE: 4 obs hidden.
```

De acuerdo a la salida anterior, se tiene que:

- El modelo empleado es significativo al 0.0141
- El r-cuadrado es 0.9998, lo cual indica que un 99.98% de los datos se ajustan al modelo
- Todos los coeficientes de modelo también son significativos: intercepto (0.0023), efecto lineal (0.0091) y efecto cuadrático (0.0091)
- Los valores observados y estimados se sobreponen en el gráfico, lo que evidencia aún más el buen ajuste del modelo.
- La ecuación despejada es: Rend = 30.20 + 0.2475 (Nit) – 0.00156 (Nit).

Para hacer el análisis de varianza, si los niveles no son igualmente espaciados, se excluye la variable Nit de CLASSES y se modifica el modelo como sigue:

```
MODEL Rend = Bloque Nit Nit*Nit Nit*Nit*Nit;
```

Se comparan las probabilidades asociadas a cada grado (lineal, cuadrático y cúbico) y se selecciona la mejor.

Comparación de medias usando contrastes ortogonales

Algunas veces, debido a la naturaleza de los tratamientos, se requiere hacer comparaciones de grupos de tratamientos, en vez de comparaciones entre promedios de tratamientos individuales.

Considérese un experimento para estudiar el efecto de fertilizantes e irrigación en el crecimiento de zanahorias en el que se usaron dos fertilizantes: sulfato de amonio (SA) y fosfato monocálcico (FM), lo mismo que un control (sin fertilizante). También se les dió a las parcelas irrigación fuerte e irrigación moderada durante el crecimiento.

Seis tratamientos fueron aplicados al azar a 12 parcelas de zanahoria en un diseño completamente al azar y se midió el promedio del peso de la zanahoria por parcela. Los datos se presentan a continuación:

	Irrigación fuerte			**Irrigación moderada**		
	SA	**FM**	**Control**	**SA**	**FM**	**Control**
	89	84	72	56	85	61
	110	89	80	54	81	40
Tratamientos	**1**	**2**	**3**	**4**	**5**	**6**

En este caso sería interesante hacer las siguientes comparaciones:

1) Irrigación fuerte contra irrigación moderada
2) SA contra FM
3) SA vs FM en irrigación fuerte y SA vs FM en irrigación moderada
4) Fertilizado contra no fertilizado (control vs SA y MP)
5) Fertilizado vs no fertilizado en irrigación fuerte y fertilizado vs no fertilizado en irrigación moderada.

Lo más apropiado en estos tratamientos, es hacer compraciones de grupos de tratamientos y no pruebas de comparación múltiple (Ej. Duncan, Tukey). Se debe usar la instrucción CONTRAST en el procedimiento GLM para hacer las pruebas de grupos de tratamientos .

El programa en SAS para analizar este experimento sería:

```
DATA Constras;
INPUT Trat Rend @@;
CARDS;
1   89  1  110  2   84  2   89  3   72  3   80
4   56  4   54  5   85  5   81  6   61  6   40
;
PROC GLM;
    CLASS Trat;
    MODEL Rend = Trat;
    CONTRAST 'IF vs IM' Trat 1  1  1  -1  -1  -1;
    CONTRAST 'SA vs FM' Trat 1 -1  0   1  -1   0;
    CONTRAST 'SA vs FM dentro irrig' Trat 1  -1  0  -1  1  0;
    CONTRAST 'Fert vs No fert' Trat 1  1  -2  1  1  -2;
    CONTRAST 'Fert vs No fert en irri' Trat 1  1  -2  -1  -1  2;
RUN;
```

Los coeficientes para cada contraste deben sumar cero. Para obtener los coeficientes debe tener en cuenta el orden en que están los tratamientos y el peso que hay que darle a cada media. Por ejemplo, para el primer contraste, los tratamientos con irrigación fuerte son los tres primeros y los tratamientos con irrigación moderada son los tres últimos (4,5,6) de acuerdo al orden en que fueron puestos en el programa. Los coeficientes serían: 1,1,1,-1,-1,-1. Esto le dice a SAS que compare la media de las medias de los tratamientos con coeficientes positivos, contra la media de las medias de los tratamientos con coeficientes negativos.

Las comparaciones hechas en este ejemplo tienen una propiedad importante. Al considerar cualquier par de contrastes, la suma de los productos de los coeficientes que ocupan igual posición es cero. Esto se cumple para cualquier par de contrastes. Las comparaciones con esta propiedad son llamadas comparaciones ortogonales.

La salida del programa SAS es la siguiente:

The GLM Procedure

Class Level Information

Class	Levels	Values
Trat	6	1 2 3 4 5 6

Number of observations 12

Dependent Variable: Rend

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	3595.416667	719.083333	8.71	0.0101
Error	6	495.500000	82.583333		
Corrected Total	11	4090.916667			

R-Square	Coeff Var	Root MSE	Rend Mean
0.878878	12.10327	9.087537	75.08333

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Trat	5	3595.416667	719.083333	8.71	0.0101

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Trat	5	3595.416667	719.083333	8.71	0.0101

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
IF vs IM	1	1800.750000	1800.750000	21.81	0.0034
SA vs FM	1	112.500000	112.500000	1.36	0.2874
SA vs FM dentro irrig	1	840.500000	840.500000	10.18	0.0188
Fert vs No fert	1	840.166667	840.166667	10.17	0.0188
Fert vs No fert en irri	1	1.500000	1.500000	0.02	0.8972

Como se ve en la salida del SAS, se encuentran diferencias significativas entre irrigación fuerte y moderada **(0.0034)**, también entre fertilizado y no fertilizado **(0.0188)**, por último, entre SA vs FM dentro de irrigaciones **(0.0188)**. Los otros dos contrastes no son significativos. En este ejemplo, los cinco grados de libertad de los tratamientos fueron divididos en cinco contrastes independientes, cada uno con un grado de libertad. La suma de las sumas de cuadrados de los contrastes debe ser igual a la suma de cuadrados de los tratamientos.

Análisis de Covarianza

Se hizo una prueba del efecto de tres tratamientos al suelo sobre crecimiento en altura de arbolitos de dos años. Los tratamientos se asignaron al azar a las tres parcelas dentro de cada uno de los 10 bloques. Cada parcela incluía 50 arbolitos. La media del crecimiento de cinco años fue el criterio para evaluar los tratamientos. Las alturas iniciales y los crecimientos de cinco años, todos ellos medidos en pies, fueron:

	Tratamiento					
	A		**B**		**C**	
Bloque	**Altura**	**Crecim**	**Altura**	**Crecim**	**Altura**	**Crecim**
1	3.6	8.9	3.1	10.7	4.7	12.4
2	4.7	10.1	4.9	14.2	2.6	9.0
3	2.6	6.3	0.8	5.9	1.5	7.4
4	5.3	14.0	4.6	12.6	4.3	10.1
5	3.1	9.6	3.9	12.5	3.3	6.8
6	1.8	6.4	1.7	9.6	3.6	10.0
7	5.8	12.3	5.5	12.8	5.8	11.9
8	3.8	10.8	2.6	8.0	2.0	7.5
9	2.4	8.0	1.1	7.5	1.6	5.2
10	5.3	12.6	4.4	8.4	4.8	10.7

Al hacer un análisis de varianza para el crecimiento, encontramos que no hay pruebas de una diferencia real en el crecimiento debido a tratamientos.

Sin embargo, existen razones para creer que, en el caso de arbolitos jóvenes, el crecimiento está afectado por la altura inicial. La posibilidad de que los efectos de los tratamientos estén encubiertos por diferencias en las alturas iniciales, plantea la duda de cómo hubieran sido los efectos de los tratamientos si la altura inicial de los árboles jóvenes hubiera sido igual.

El análisis de covarianza usando la variable altura inicial como covariable, ajustaría los datos de tal manera que se probarían los tratamientos, como si todas las alturas iniciales fueran las mismas.

El siguiente programa SAS hace un análisis de covarianza para los datos anteriores:

```
DATA Covar;
INPUT Bloque Trat $ Altura Crecim;
CARDS;
 1   A   3.6   8.9
 2   A   4.7  10.1
(resto de los datos)
 ;
PROC GLM;
    CLASS Bloque Trat;
    MODEL Crecim = Bloque Trat Altura;
RUN;
```

En la instrucción MODEL, se incluyó la variable Altura. Debido a que Altura no está definida como una clase (en la instrucción CLASS), SAS la considera como una covariable. Si se tienen más covariables en el experimento, éstas se incluyen a la derecha del signo "=", después de las fuentes de variación (en este ejemplo Bloque y Tratamiento). La salida del programa anterior es la siguiente:

```
                          The GLM Procedure

                     Class Level Information

              Class          Levels    Values

              Bloque             10    1 2 3 4 5 6 7 8 9 10

              Trat                3    A B C

                      Number of observations    30

Dependent Variable: Crecim

                                                Sum of
Source                     DF          Squares      Mean Square    F Value    Pr > F

Model                      12      149.7823876       12.4818656       6.09    0.0004

Error                      17       34.8296124        2.0488007

Corrected Total            29      184.6120000

            R-Square     Coeff Var      Root MSE    Crecim Mean

            0.811336      14.69572      1.431363       9.740000
```

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Bloque	9	110.3786667	12.2642963	5.99	0.0008
Trat	2	6.6560000	3.3280000	1.62	0.2262
Altura	1	32.7477209	32.7477209	15.98	0.0009

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Bloque	9	10.70245381	1.18916153	0.58	0.7954
Trat	2	11.68343057	5.84171529	2.85	0.0855
Altura	1	32.74772090	32.74772090	15.98	0.0009

De la salida, se puede ver que la variable altura inicial explica mucho de la variación total y es altamente significativa **(0.0009)**. Para ver si los tratamientos son significativamente diferentes se examinan las sumas de cuadrados tipo III y se ve que el valor de F ajustado para tratamientos es **2.85**. Este valor es mucho más alto que el F sin ajustar **(1.62)**. Se puede entonces ver que el análisis de covarianza sirvió para aumentar la F, y en este caso, se podrían aceptar diferencias entre tratamientos con un alfa = **0.0855**.

Si se desean calcular las medias ajustadas de cada tratamiento y al mismo tiempo ver cuáles medias son diferentes, se agrega la siguiente línea al programa:

```
LSMEANS Trat / PDIFF;
```

Esta línea produce la siguiente salida:

The GLM Procedure
Least Squares Means

Trat	Crecim LSMEAN	LSMEAN Number
A	9.3142570	1
B	10.6534498	2
C	9.2522932	3

Least Squares Means for effect trat
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: Crecim

i/j	1	2	3
1		0.0687	0.9270
2	0.0687		0.0440
3	0.9270	0.0440	

NOTE: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used.

Las medias ajustadas de cada tratamiento aparecen debajo de la columna Crecim/LSMEAN. Abajo, aparece una matriz de probabilidades apoyando la hipótesis nula; Ho: media ajustada i = media ajustada j. En este ejemplo la media del tratamiento A es diferente a la del tratamiento B (la probabilidad obtenida, **0.0687**, es menor que 0.10. También la media B es diferente a la media C **(0.0440)**, mientras que la media A y C no son diferentes **(0.9270)**.

Supuestos del análisis de covarianza

A continuación se presentan los supuestos que se deben cumplir para que el análisis de covarianza sea un análisis válido:

- Los valores x (covariable), se miden sin error, e independientes de los tratamientos
- Homogeneidad de las regresiones β_i dentro de los grupos
- La regresión utilizada es lineal e independiente de los tratamientos. Los tratamientos no influyen en la covariable x.
- Los residuos se distribuyen normalmente e independientemente con media 0 y varianza común. Esto implica que las regresiones poblacionales de los grupos tienen la misma pendiente.
- Existe correlación entre la variable dependiente y la covariable. Si no hay correlación, el análisis de covarianza no tendría ventajas sobre el análisis de varianza.

Tipos de sumas de cuadrados calculados por el GLM

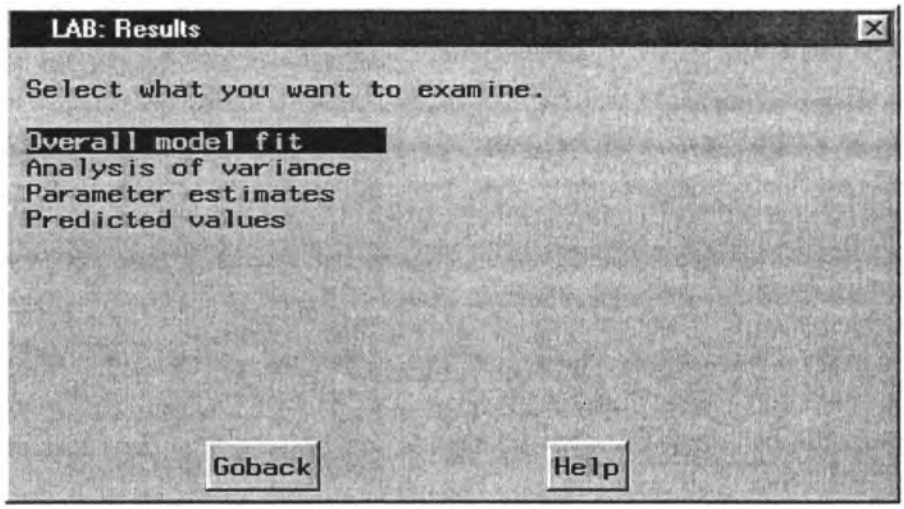
El Proc GLM, produce cuatro diferentes tipos de sumas de cuadrados relacionadas con diferentes interpretaciones teóricas. Estos cuatro tipos de sumas de cuadrados son llamados Tipo I, Tipo II, Tipo III y Tipo IV. Para nuestros propósitos, sólo interesan las sumas de cuadrados Tipo I y Tipo III.

Las sumas de cuadrados Tipo I corresponden a añadir cada fuente (factor) secuencialmente al modelo en el orden dado. Este tipo de sumas de cuadrados no es útil para una estructura de varias entradas con datos no balanceados, pero sirve para diseños ortogonales (p. ej. bloques al azar, parcelas divididas, etc. sin datos perdidos), diseños anidados y modelos polinomiales.

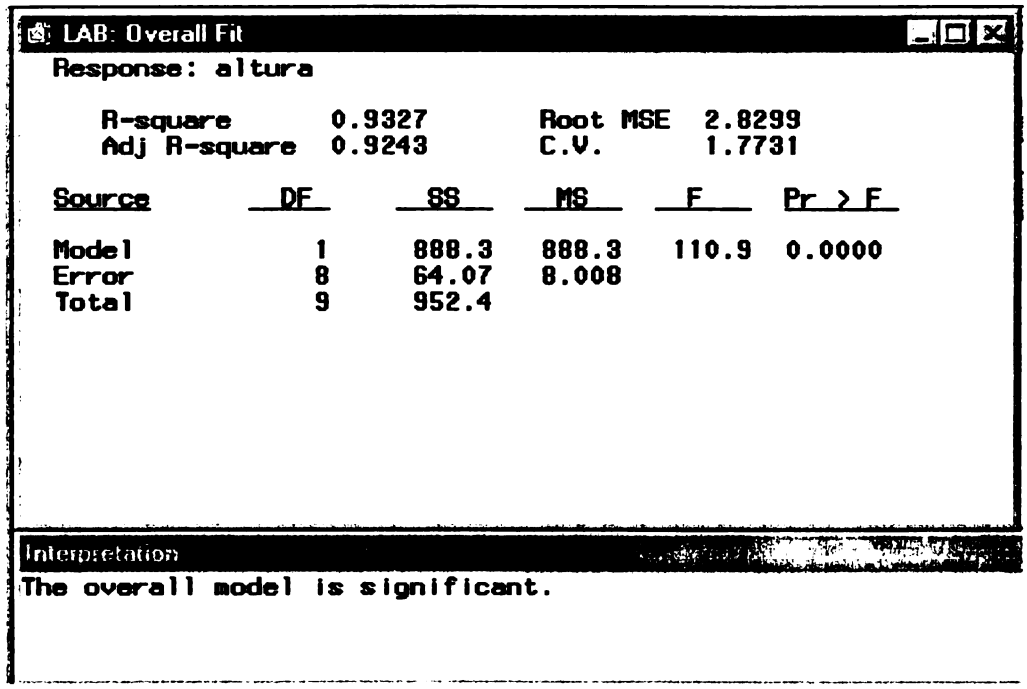
Las sumas de cuadrados Tipo III son sumas de cuadrados parciales. Su uso principal es en situaciones que requieren una comparación de efectos principales, aún bajo la presencia de interacción. Cada efecto es ajustado después de los otros efectos. Deben usarse cuando se tienen datos no balanceados y en análisis de covarianza, pero no en diseños o situaciones donde el orden

supuestos que se violan, se da click sobre el botón **Assumptions** y despliega la ventana con los supuestos que puede verificar.

Para obtener otros resultados, se da click sobre el botón Results y se despliega la siguiente ventana:

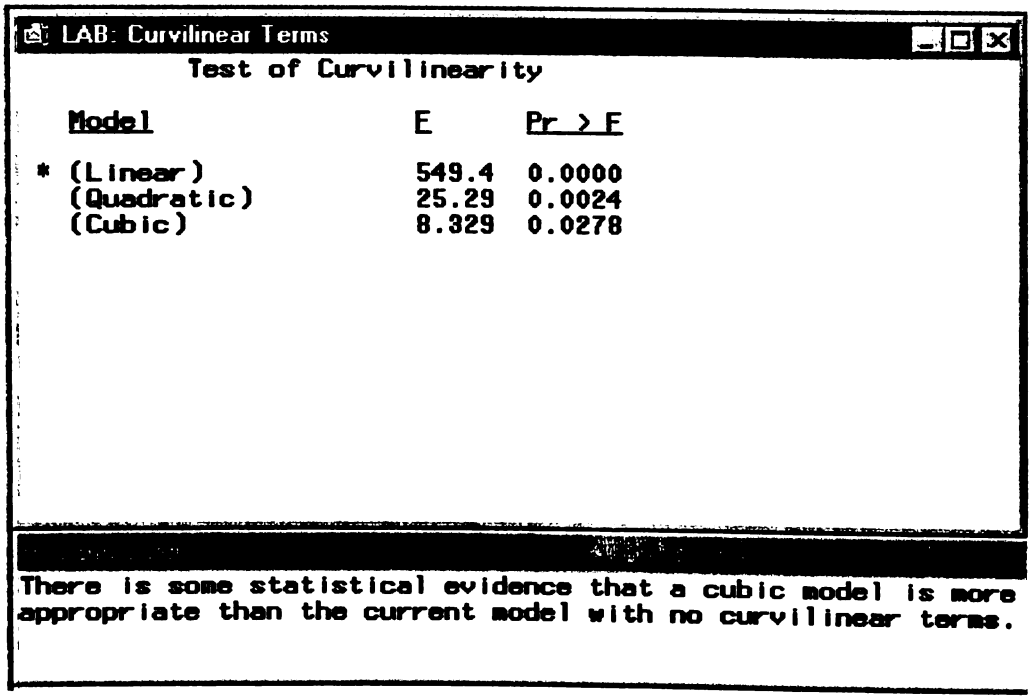


De click sobre el resultado que quiere obtener. En este caso se va a obtener el análisis de varianza del modelo de regresión y otras estadísticas. Para esto, de click sobre la opción deseada (**Overall model fit**): y se despliega la siguiente ventana:



La interpretación indica que el modelo global es significativo.

Para ver si hay un modelo que se pueda ajustar mejor, de click sobre el botón **Assumptions** y de click sobre la opción **Curvilinearity**. Se despliega la siguiente ventana:



Como puede ver en la interpretación, el SAS dice que existe alguna evidencia estadística que un modelo cúbico es más apropiado que el modelo actual.

de los factores no puede ser cambiado (diseño de parcelas divididas, ajustes de modelos polinomiales).

Supóngase que se ajusta el siguiente modelo:

```
MODEL Y = A B A*B;
```

Las sumas de cuadrados Tipo I y Tipo III ajustarían los factores de la siguiente manera:

Efecto	Tipo I	Tipo III
A	$R(\alpha \mid \mu)$	$R(\alpha \mid \mu, \beta, \alpha\beta)$
B	$R(\beta \mid \mu, \alpha)$	$R(\beta \mid \mu, \alpha, \alpha\beta)$
A*B	$R(\alpha\beta \mid \mu, \alpha, \beta)$	$R(\alpha\beta \mid \mu, \alpha, \beta)$

Como se puede ver la sumas de cuadrado Tipo I ajustan los efectos secuencialmente, primero el factor A, luego el B y por último la interacción. Las sumas de cuadrados Tipo III ajustan cada factor después de haber ajustado los otros factores.

Componentes de Varianza

Cuando el efecto que se estudia es aleatorio, es decir, se han seleccionado los niveles del factor de interés de una población grande de niveles, lo que interesa en este caso, es ver el aporte de las varianzas dentro y entre las unidades experimentales. Para el análisis de efectos aleatorios se habla comúnmente de las fuentes de variabilidad 'entre' y 'dentro' los grupos muestras, para efectos fijos se prefiere los términos 'modelo' y 'error'. El procedimiento en SAS para obtener los componenetes de varianza es el VARCOMP. A continuación un ejemplo utilizando este procedimiento:

Un beneficio recibe café de 150 productores. Un aspecto crítico en la exportación del producto es la calidad. Para que el café cumpla los requerimientos de calidad y pueda comercializarse, debe existir una variabilidad pequeña en la alta calidad, para que sea rentable para el beneficio. El beneficio seleccionó aleatoriamente a tres productores y obtuvo 10 muestras de granos de café para cada productor y midió el índice de calidad.

Lo que interesa en este caso es, cuanto contribuye el factor 'productor' a la variabilidad total (componente de varianza). Lo que interesa es si hay diferencias significativas entre los productores, ya que se seleccionaron al azar.

A continuación se presentan los datos de las muestras de café para cada productor.

Muestra	Prod1	Prod2	Prod3	Muestra	Prod1	Prod2	Prod3
1	88.0	85.9	94.2	6	89.0	86.0	93.8
2	88.0	88.6	91.5	7	86.0	91.0	92.5
3	94.8	90.0	92.0	8	92.9	89.6	93.2
4	90.0	87.1	96.5	9	89.0	93.0	96.2
5	93.0	85.6	95.6	10	93.0	87.5	92.5

El siguiente programa SAS hace un análisis de componentes de varianza dentro y entre productores:

```
Options ls=80 ps=60 pageno=1 nodate;
Data Cafe;
Do Productor=1 to 3;
   Do Muestra=1 to 10;
      Input Indice @@;
      Output;
   End;
End;
Cards;
88.0 88.0 94.8 90.0 93.0 89.0 86.0 92.9 89.0 93.0
85.9 88.6 90.0 87.1 85.6 86.0 91.0 89.6 93.0 87.5
94.2 91.5 92.0 96.5 95.6 93.8 92.5 93.2 96.2 92.5
;
Proc Varcomp;
Title 'Componentes de Varianza';
Classes Productor;
   Model Indice = Productor;
Run;
```

La salida del programa anterior se muestra a continuación:

```
                    Componentes de Varianza
          Variance Components Estimation Procedure

                    Class Level Information
              Class          Levels    Values
              Productor           3    1 2 3

              Variance Component         Indice
              Var(Productor)             6.81237 (1)
              Var(Error)                 5.81859 (2)
```

De acuerdo a la salida de SAS, en lo referente a la composición de varianzas, la varianza correspondiente a Productores es 6.81237 **(1)** y la varianza dentro de Productores es 5.81859 **(2)**.

Como se puede ver, en ambos casos, las varianzas son mayores que cero. Analizando estas varianzas en términos relativos, se obtienen los siguientes resultados:

Varianza Total = 6.81237 + 5.81859 = 12.63096
Varianza Entre Productores = 6.81237 / 12.63096 = 54%
Varianza Dentro Productores = 5.81859 / 12.63096 = 46%

De acuerdo a los resultados relativos, se ve que la variabilidad entre productores es grande y comprende mas de la mitad de la variancia total.

Se concluye que, el "productor" es una fuente importante en la determinación de la variabilidad del producto, y que el beneficio debe tomar medidas para mejorar la homogeneidad de la calidad de la materia prima que entregan los productores.

Estimaciones de componentes de varianzas negativas

Por definición, un componente de varianza es positivo. Sin embargo, existen casos en que los componentes de varianza son negativos. Algunas razones por lo que esto puede suceder son:

- La variabilidad en los datos puede ser tan grande que produce estimaciones negativas, aunque lo normal sea que los componentes de varianza sean positivos
- Los datos pueden contener 'outliers' (valores extremos). Para detectar esto, se pueden utilizar técnicas gráficas para identificarlos.
- El modelo que se está utilizando, no es el más apropiado. Bajo algunos modelos estadísticos para análisis de componentes de varianza, estimaciones negativas son indicación que las observaciones en sus datos están negativamente correlacionadas.

¿Cómo proceder ante estimaciones de componentes de varianzas negativas?

A continuación se presentan algunas acciones a seguir cuando las estimaciones de componentes de varianzas son negativas:

- Aceptar la estimación como evidencia de un valor cero verdadero, y utilizar cero como la estimación, sin embargo, se debe considerar que el estimador no será imparcial.
- Aceptar la estimación negativa, reconociendo que para cálculos subsecuentes, los resultados pueden no tener sentido.
- Interpretar la estimación negativa como indicación de un modelo estadístico incorrecto.
- Utilizar un modelo estadístico diferente para la estimación de la varianza

Recolectar más datos y analizarlos separadamente o en conjunto con los existentes y verificar si la estimación es positiva.

Un ejemplo donde se presenta estimaciones de componentes de varianzas negativas

Se realizó un experimento para comparar el efecto del espaciamiento de cinco surcos sobre el rendimiento de dos variedades de soya. El diseño era de parcelas divididas, con variedades como tratamientos de parcelas completas, en un diseño de bloques completos al azar; los espaciamientos entre surcos se aplicaron a subparcelas. A continuación el programa SAS que analiza los datos de este experimento:

```
data dividida;
do Variedad = 1 to 3;
   do Espacio = 18 to 42 by 6;
      do Block = 1 to 6;
         input Rend @@;
         output;
       end;
   end;
end;
cards;
33.6  37.1  34.1  34.6  35.4  36.1
31.1  34.5  30.5  32.7  30.7  30.3
33.0  29.5  29.2  30.7  30.7  27.9
28.4  29.9  31.6  32.3  28.1  26.9
31.4  28.3  28.9  28.6  18.5  33.4
28.0  25.5  28.3  29.4  27.3  28.3
23.7  26.2  27.0  25.8  26.8  23.8
23.5  26.8  24.9  25.3  21.4  22.0
25.0  25.3  25.6  26.4  24.6  24.5
25.7  23.2  23.4  25.6  24.5  22.9
25.3  24.5  26.8  34.6  22.1  26.4
28.9  34.2  28.6  26.5  32.5  24.5
32.3  27.6  32.7  28.9  29.6  32.8
28.5  25.8  33.2  30.8  24.8  29.7
29.4  35.1  27.9  33.5  27.5  22.4
;
PROC GLM;
  CLASS Block Variedad Espacio;
  MODEL Rend = Block Variedad Variedad*Block Espacio Variedad*Espacio;
  RANDOM Variedad*Block / TEST;
  TEST H=Variedad E=Variedad*Block;
RUN;
PROC VARCOMP;
  CLASS Block Variedad Espacio;
  MODEL Rend = Block Variedad Espacio Variedad*Espacio Variedad*Block
/FIXED=4;
RUN;
PROC MIXED;
  CLASS Block Variedad Espacio;
  MODEL Rend = Block Variedad Espacio Variedad*Espacio;
  RANDOM Variedad*Block;
RUN;
```

Las siguientes son las salidas más relevantes del programa:

```
                          The GLM Procedure

                       Class Level Information

                Class         Levels    Values
                Block              6    1 2 3 4 5 6
                Variedad           3    1 2 3
                Espacio            5    18 24 30 36 42
                     Number of observations    90

Dependent Variable: Rend
                                    Sum of
Source                      DF       Squares    Mean Square   F Value   Pr > F
Model                       29     857.361222     29.564180      3.74   <.0001
Error                       60     474.568667      7.909478
Corrected Total             89    1331.929889

            R-Square     Coeff Var      Root MSE     Rend Mean
            0.643698      9.902357      2.812379      28.40111

Source                      DF    Type III SS    Mean Square   F Value   Pr > F

Block                        5     76.9018889     15.3803778      1.94   0.1002
Variedad                     2    478.2948889    239.1474444     30.24   <.0001
Block*Variedad              10     25.4211111      2.5421111      0.32   0.9724
Espacio                      4     71.0915556     17.7728889      2.25   0.0745 (1)
Variedad*Espacio             8    205.6517778     25.7064722      3.25   0.0039 (2)

Tests of Hypotheses Using the Type III MS for Block*Variedad as an Error Term

Source                      DF    Type III SS    Mean Square   F Value   Pr > F
Variedad                     2    478.2948889    239.1474444     94.07   <.0001 (3)

                 Variance Components Estimation Procedure

                       Class Level Information
                Class         Levels    Values
                Block              6    1 2 3 4 5 6
                Variedad           3    1 2 3
                Espacio            5    18 24 30 36 42
                     Number of observations    90

                         MIVQUE(0) Estimates

                   Variance Component          Rend
                   Var(Block*Variedad)     -1.07347 (4)
                   Var(Error)               7.90948

                         The Mixed Procedure

                       Class Level Information
            Class       Levels    Values
            Block            6    1 2 3 4 5 6
            Variedad         3    1 2 3
            Espacio          5    18 24 30 36 42
```

Covariance Parameter Estimates

Cov Parm	Estimate
Block*Variedad	0 (5)
Residual	7.1427

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Block	5	10	2.15	0.1414
Variedad	2	10	33.48	<.0001 (6)
Espacio	4	60	2.49	0.0527 (7)
Variedad*Espacio	8	60	3.60	0.0018 (8)

De acuerdo a la salida se ve que los resultados obtenidos por el GLM y MIXED son diferentes. El valor de F para Variedades calculado por GLM es 94.07 (3), mientras que el valor de F calculado por MIXED es 33.48 (6). Para Espacio, los valores de F son 2.25 (1) para GLM y 2.49 para MIXED.

Esta situación se presenta debido a estimaciones de componentes de varianzas negativas. En el programa se realiza el VARCOMP, con el propósito de determinar esta situación. El componente de varianza estimado para Block*Variedad es –1.07347 (4). Como la varianza es negativa, el MIXED asume cero como valor verdadero para este estimación.(5)

Las diferencias entre ambos procedimientos se debe principalmente al tipo de método que utiliza el MIXED. Si usted agrega al programa anterior, la opción **METHOD=TYPE3** en la instrucción **PROC MIXED**, los resultados de ambos procedimientos serán los mismos.

Experimentos con repeticiones de mediciones

En algunos experimentos, se realiza más de una medición en los mismos objetos/sujetos (unidad experimental). Algunos ejemplos de estos experimentos son:

- Mediciones de valores fisiológicos en los mismos animales en la mañana, tarde y noche.
- Medir el desempeño de trabajadores antes y después de un entrenamiento
- Medir el incremento radial de árboles en tres diferentes alturas
- Medir el contenido de un nutriente en el suelo una, dos y tres semanas después de haber fertilizado

Para este tipo de experimentos, normalmente solo interesa la variabilidad dentro de los objetos/sujetos, ya que es ahí donde se ven los efectos de los tratamientos.

Uno de los supuestos del ANDEVA es que las unidades experimentales son independientes, lo que se garantiza a través de la aleatorización. En el caso de mediciones repetidas, no se aleatoriza

independientemente la asignación de los tratamientos (que se aplica repetidamente), si no se aleatoriza solamente la selección de los objetos. Esto significa, que se viola uno de los supuestos del ANDEVA univariado.

Sin embargo, se ha encontrado, que sí se puede aplicar un análisis univariado (por ejemplo, como un diseño en parcelas divididas), siempre y cuando la estructura de la matriz de varianza y covarianza cumple con ciertas condiciones; una de estas es que las correlaciones entre todos los pares de grupos son iguales. Otra condición menos fuerte es la llamada condición Huynh – Feldt que se prueban con la prueba de Mauchly (véase salida de SAS más adelante).

Si estas condiciones no se cumplen, es mejor considerar las respuestas de las mediciones repetidas, como diferentes variables, y aplicar un MANOVA (Análisis de varianza multivariado).

A continuación se presenta un experimento con mediciones repetidas:

		Mediciones de la altura (cm)		
Planta	**Tratamiento**	**1**	**2**	**3**
1	1	7.0	13.0	21.0
2	1	8.0	15.0	22.0
3	1	6.5	12.0	23.0
4	2	7.5	14.0	23.0
5	2	9.0	16.0	28.0
6	2	7.0	15.0	23.0
7	3	5.0	7.0	13.0
8	3	6.5	12.0	17.0
9	3	7.5	14.0	22.0

Seguidamente, el programa SAS para analizar los datos anteriores, utilizando la técnica MANOVA. En este experimento, se tienen 9 plantas, a las cuales se asignó al azar uno de 3 tratamientos (Factor 'TRAT'). De cada planta, se efectuaron 3 mediciones (Factor 'MED'), en diferentes puntos en el tiempo. Obviamente , la asignación de los diferentes niveles de MED no es aleatoria.

```
OPTIONS LS=78;
Data RepMed;
Input Planta Trat Med1-Med3 @@;
Cards;
1 1 7.0 13.0 21.0 2 1 8.0 15.0 22.0
3 1 6.5 12.0 23.0 4 2 7.5 14.0 23.0
5 2 9.0 16.0 28.0 6 2 7.0 15.0 23.0
7 3 5.0  7.0 13.0 8 3 6.5 12.0 17.0
9 3 7.5 14.0 22.0
;
PROC GLM;
   Classes Trat;
   Model Med1-Med3=Trat;
   REPEATED Med 3 / PRINTE;
run;
```

A continuación la salida del programa SAS anterior (en donde viene primero el análisis de varianza univariado para cada uno de los niveles del factor medición Med1, Med2, Med3; esta parte no es incluída aquí).

The GLM Procedure

Repeated Measures Analysis of Variance
Repeated Measures Level Information

Dependent Variable	Med1	Med2	Med3
Level of Med	1	2	3

Partial Correlation Coefficients from the Error SSCP Matrix / Prob > |r|

DF = 6	Med1	Med2	Med3
Med1	1.000000	0.880705	0.840191
		0.0088	0.0180
Med2	0.880705	1.000000	0.794998
	0.0088		0.0326
Med3	0.840191	0.794998	1.000000
	0.0180	0.0326	

E = Error SSCP Matrix
Med_N represents the contrast between the nth level of Med and the last

	Med_1	Med_2
Med_1	32.833	20.667
Med_2	20.667	22.000

Partial Correlation Coefficients from the Error SSCP Matrix of the Variables Defined by the Specified Transformation / Prob > |r|

DF = 6	Med_1	Med_2
Med_1	1.000000	0.768956
		0.0433
Med_2	0.768956	1.000000
	0.0433	

The GLM Procedure
Repeated Measures Analysis of Variance

Sphericity Tests

Variables	DF	Mauchly's Criterion	Chi-Square	Pr > ChiSq
Transformed Variates	2	0.3927532	4.6728692	0.0967
Orthogonal Components	2	0.7586913	1.3808018	0.5014 (1)

Manova Test Criteria and Exact F Statistics
for the Hypothesis of no Med Effect (2)
H = Type III SSCP Matrix for Med
E = Error SSCP Matrix

S=1 M=0 N=1.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.01755058	139.95	2	5	<.0001
Pillai's Trace	0.98244942	139.95	2	5	<.0001
Hotelling-Lawley Trace	55.97817087	139.95	2	5	<.0001
Roy's Greatest Root	55.97817087	139.95	2	5	<.0001

Manova Test Criteria and F Approximations
for the Hypothesis of no Med*Trat Effect
H = Type III SSCP Matrix for Med*Trat (3)
E = Error SSCP Matrix

S=2 M=-0.5 N=1.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.37969347	1.56	4	10	0.2593
Pillai's Trace	0.62298596	1.36	4	12	0.3056
Hotelling-Lawley Trace	1.62664659	2.00	4	5.1429	0.2303
Roy's Greatest Root	1.62229669	4.87	2	6	0.0555

NOTE: F Statistic for Roy's Greatest Root is an upper bound.
NOTE: F Statistic for Wilks' Lambda is exact.

The GLM Procedure
Repeated Measures Analysis of Variance
Tests of Hypotheses for Between Subjects Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Trat	2	83.68518519	41.84259259	3.32	0.1072 (4)
Error	6	75.72222222	12.62037037		

Repeated Measures Analysis of Variance
Univariate Tests of Hypotheses for Within Subject Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Med	2	917.6296296	458.8148148	241.72	<.0001 (5)
Med*Trat	4	26.5925926	6.6481481	3.50	0.0408
Error(Med)	12	22.7777778	1.8981481		

La prueba de Mauchly, la cual indica aquí (Pr > ChiSq: 0.5014 **(1)**), que un Andeva univariado (como parcelas divididas) sería probablemente aceptable.

El análisis multivariado indica un efecto muy significativo de Med **(2)**, y no efecto de la interacción Med*Trat **(3)**. Para este análisis, como puede ver, el SAS presenta varios estadísticos: Wilks' Lambda, Pillai's Trace, entre otros.

Finalmente, viene impreso el resultado del ANDEVA univariado. Para el factor TRAT **(4)**, este es correcto (según el esquema de aleatorización explicado).

Para MED, el resultado es igualmente significativo, sin embargo con un valor de F mucho más alto (241.72 (5), en vez de 139.95 (2)). La interacción es significativa (para un α = 5%) en el Andeva univariado, a pesar de que el Andeva correcto multivariado indica que no!. Con la capacidad de las computadoras hoy día, siempre es recomendable aplicar el ANDEVA multivariado, aunque las pruebas indiquen que se pueda realizar el ANDEVA univariado.

Experimentos con efectos mixtos (PROC MIXED)

Algunos experimentos presentan efectos mixtos, es decir, el factor A puede ser fijo, el factor B aleatorio u otras posibles combinaciones.Para diseños con 2 y más factores, las pruebas de los efectos hay que hacerlas tomando en cuenta el carácter de los efectos. A continuación se presenta un cuadro,el cual muestra la varianza correcta en el denominador de la prueba de F, para las combinaciones de efectos en un diseño con dos factores:

Varianza en el denominador de la prueba de F			
Varianza para probar (Numerador)	Modelo I A fijo B fijo	Modelo II A fijo B aleatorio	Modelo III A aleatorio B aleatorio
σ^2_A	σ^2_{Error}	σ^2_{AxB}	σ^2_{AxB}
σ^2_B	σ^2_{Error}	σ^2_{Error}	σ^2_{AxB}
σ^2_{AxB}	σ^2_{Error}	σ^2_{Error}	σ^2_{Error}

El procedimiento MIXED del SAS, permite analizar experimentos en donde hay factores de efectos aleatorios. Otra aplicación es para diseños en parcelas divididas, en donde el término de error de la parcela grande es un factor de efecto aleatorio.

El siguiente programa SAS, realiza el análisis de varianza para un diseño en parcelas divididas, utilizando el procedimiento MIXED y GLM. La diferencia en la sintáxis del PROC GLM está sobre todo en la espicificación del modelo. En MIXED, solamente los efectos fijos están en la instrucción MODEL; los efectos aleatorios se especifican en la instrucción RANDOM, luego el PROC MIXED usa los términos correctos de los errores al efectuar pruebas de F, tanto, para las pruebas de factores, como para pruebas de contrastes.

```
Data Sp;
Input Block A B Y @@;
Cards;
   1 1 1  56  1 1 2  41  1 2 1  50  1 2 2  36
   1 3 1  39  1 3 2  35  2 1 1  30  2 1 2  25
   2 2 1  36  2 2 2  28  2 3 1  33  2 3 2  30
   3 1 1  32  3 1 2  24  3 2 1  31  3 2 2  27
   3 3 1  15  3 3 2  19  4 1 1  30  4 1 2  25
   4 2 1  35  4 2 2  30  4 3 1  17  4 3 2  18
;
Proc Mixed;
Title 'Análisis con MIXED';
     Class A B Block;
     Model Y = A B A*B;
     Random Block A*Block;
Proc Glm;
Title 'Análisis con GLM';
     Classes Block A B;
      Model Y = Block A Block*A B A*B;
      Test H=A E=Block*A;
Run;
```

A continuación se presentan las salidas más relevantes del programa anterior:

Análisis con MIXED

The Mixed Procedure

Class Level Information

Class	Levels	Values
Block	4	1 2 3 4
A	3	1 2 3
B	2	1 2

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
A	2	6	4.07	0.0764 (1)
B	1	9	19.39	0.0017 (2)
A*B	2	9	4.02	0.0566 (3)

Análisis con GLM

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	2067.583333	147.684524	15.78	0.0001
Error	9	84.250000	9.361111		
Corrected Total	23	2151.833333			

R-Square	Coeff Var	Root MSE	Y Mean
0.960847	9.896259	3.059593	30.91667

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Block	3	1243.500000	414.500000	44.28	<.0001
A	2	326.583333	163.291667	17.44	0.0008
Block*A	6	240.750000	40.125000	4.29	0.0256
B	1	181.500000	181.500000	19.39	0.0017 (4)
A*B	2	75.250000	37.625000	4.02	0.0566 (5)

Tests of Hypotheses Using the Type III MS for Block*A as an Error Term

Source	DF	Type III SS	Mean Square	F Value	Pr > F
A	2	326.5833333	163.2916667	4.07	0.076 (6)

Comparando las salidas de ambos procedimientos (GLM y MIXED), los resultados obtenidos son los mismos. Para el efecto de A, el valor de F = 4.07 y la Pr > F = 0.076 **(1)** y **(6)**. Para el efecto de B, el valor de F = 19.39 y la Pr > F = 0.0017 **(2)** y **(4)**. Para el efecto de A*B, el valor de F = 4.02 y la Pr > F = 0.0566 **(3)** y **(5)**.

Un ejemplo de experimentos con efectos mixtos

Se tomaron las alturas de 18 individuos en pulgadas. Cada individuo se clasificó de acuerdo a Familia y Género. Dependiendo de la forma en que se realizó puede ser, por ejemplo:

- Con ambos efectos como fijos o,
- Con el efecto Familia como aleatorio y Género como fijo

A continuación se presenta el programa para analizar los datos:

```
Options Ls=78;
Data Alturas;
Input Familia Genero$ Altura @@;
cards;
1 F 67  1 F 66  1 F 64  1 M 71  1 M 72  2 F 63  2 F 63  2 F 67 4 M 69
2 M 69  2 M 68  2 M 70  3 F 63  3 M 64  4 F 67  4 F 66  4 M 67 4 M 67
;
Proc Mixed;
Title 'Análisis con ambos factores fijos';
   Class Familia Genero;
   Model Altura = Genero Familia Familia*Genero;
Proc Mixed;
Title 'Análisis con Familia como factor aleatorio';
   Class Familia Genero;
   Model Altura = Genero;
   Random Familia Familia*Genero;
Run;
```

Seguidamente, un resumen de la salida del programa anterior. Se muestran únicamente las Pr > F para los factores:

Análisis considerando ambos factores fijos

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Genero	1	10	17.63	0.0018 (1)
Familia	3	10	5.90	0.0139
Familia*Genero	3	10	2.89	0.0889

Análisis considerando Familia como factor aleatorio

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Genero	1	3	7.95	0.0667 (2)

Como se puede ver en la salida, la Pr > F para Género ((1) y (2)), es diferente cuando se declara a Familia como factor aleatorio. En el primer diseño, los tratamientos tienen efecto significativo, mientras que en el segundo se concluye que el componente de variancia debido a género no es significativamente diferente de cero ($\alpha = 0.05$) El análisis correcto será el que considere los efectos tal como han sido seleccionados.

Práctica

La siguiente tabla muestra la densidad de semillas de maíz en invernaderos, infectados con *Diplodia spp*, en tratamientos con varios fungicidas

	TRATAMIENTOS							
BLOQUE	A	B	C	D	E	F	G	H
1	8	16	14	10	8	8	7	12
2	8	19	16	11	7	8	6	19
3	9	24	14	12	1	3	6	9
4	7	22	13	8	1	3	6	11
5	7	19	14	7	3	3	4	9
6	5	19	13	3	2	7	4	5

Existen grupos de tratamientos, los cuales se indican a continuación:

Grupo	DESCRIPCION
A	Control sin tratamiento
B y C	Fungicidas mercúricos
D y H	Fungicidas no mercúricos, compañía I
E, F y G	Fungicidas no mercúricos, compañía II, donde F y G son fromulaciones nuevas de E

Con la información anterior, realice un análisis de varianza y haga comparaciones de grupos de tratamientos. En este caso, podrá realizar como máximo 7 comparaciones.

Concluya con base a los resultados.

CAPITULO VI: Otros componentes de SAS

El Asistente de SAS: análisis de datos interactivamente

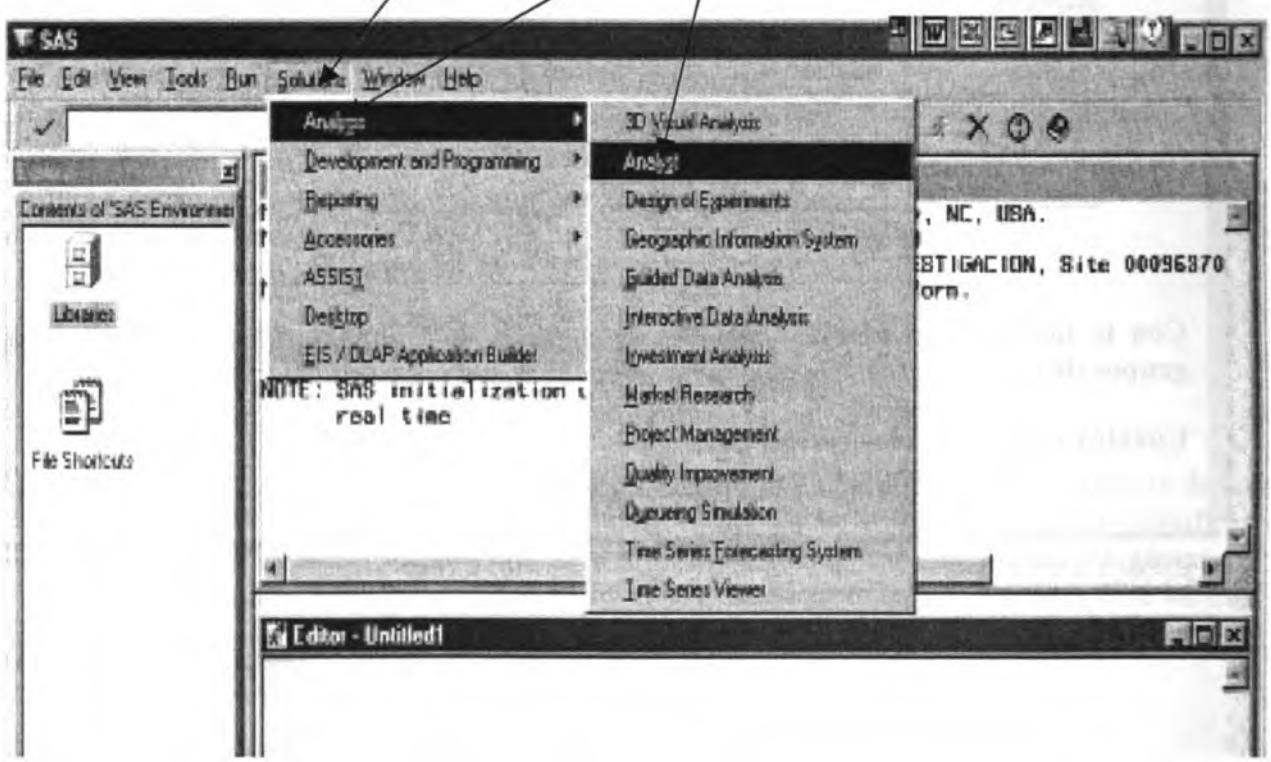
Para aquellos usuarios que no dominan la programación SAS y que vayan a realizar análisis de datos no muy complejos, el SAS incluye el **Analysis**, el cual incluye una serie de herramientas que le permiten al usuario, realizar análisis de datos interactivamente. También permite hacer gráficos fácilmente de muy buena calidad. En este documento se explicarán tres opciones de análisis de datos de esta herramienta: **Analyst, Guided Data Analysis** y el **Interactive Data Analysis.**

Analyst

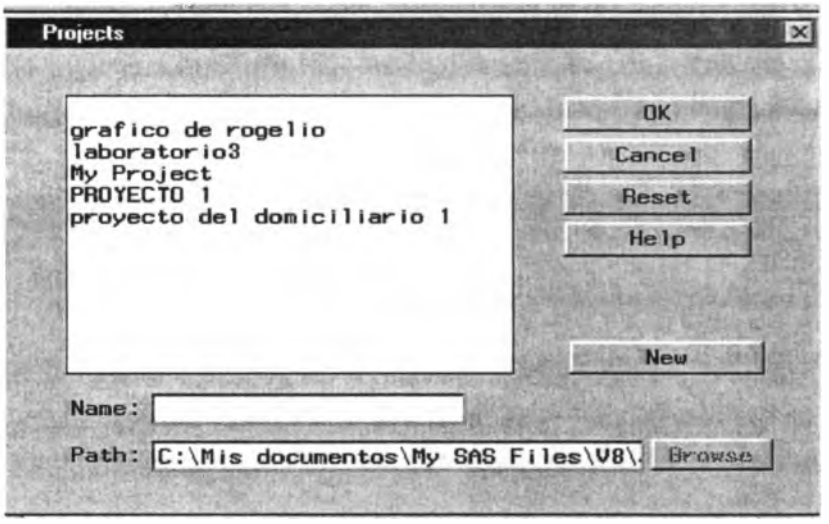
Esta herramienta se compone de un menú y una hoja similar a Excel, en la cual el usuario introduce los datos. El menú le permite acceder a los diferentes procedimientos para análisis de datos, así como a las diferentes opciones para generar gráficos. Se puede realizar análisis de varianza, estimación de regresiones, análisis multivariado, entre otros.

Para entrar al Analyst, siga los siguientes pasos:

- De click en la opción de menú **Solutions**
- En el siguiente menú, de click sobre la opción **Analysis**
- Seguidamente, de click sobre la opción **Analyst**

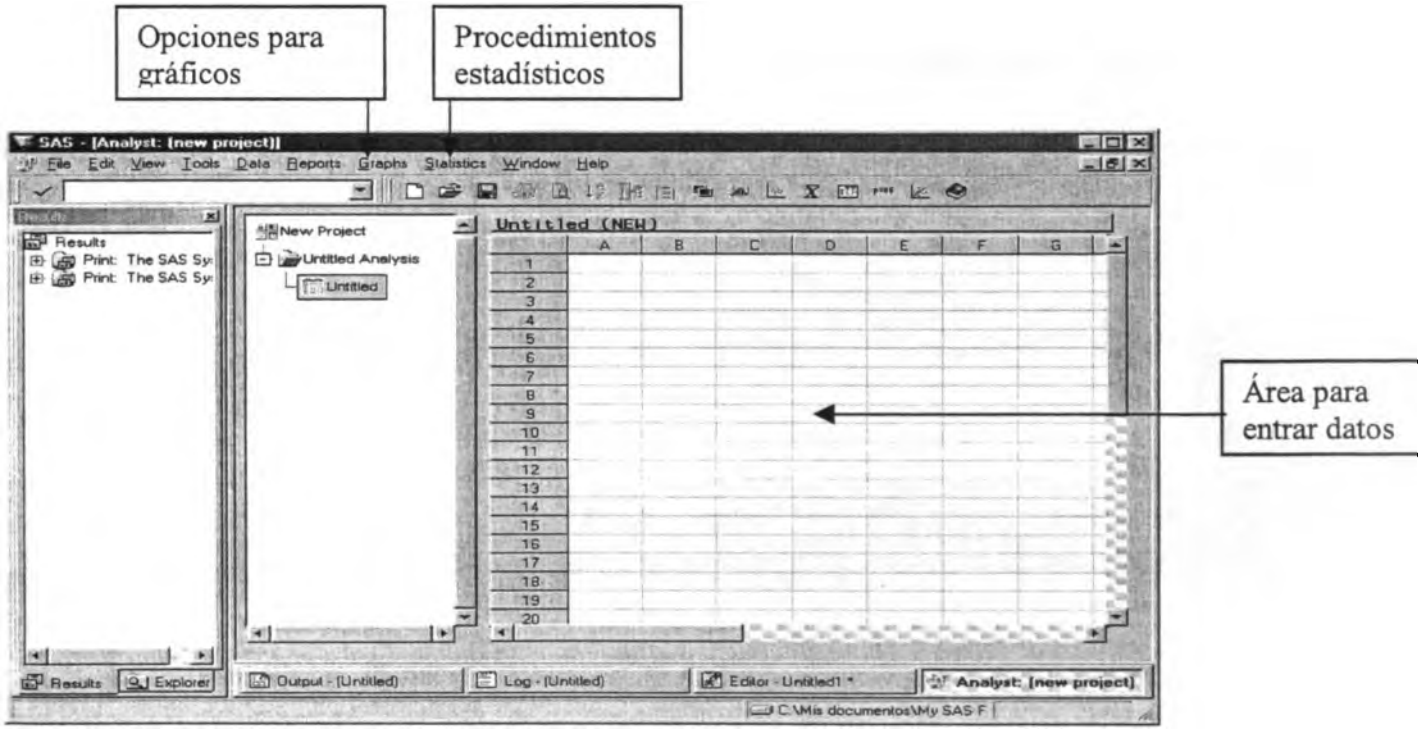


La siguiente ventana que se despliega es para seleccionar un proyecto, o bien, para crear un nuevo proyecto. Esta es una de las características del Analyst que no tienen las otras herramientas del Analysis para procesar datos. Como puede ver, la herramienta permite grabar el proyecto, el cual incluye los datos y los diferentes análisis y gráficos que se realizan. En otra sesión posterior, el usuario puede retomar el proyecto y hacer nuevos procesos a sus datos.



Una vez seleccionada la acción requerida, se despliega la hoja para la entrada de datos y el menú para análisis y gráficos.

Proceda a introducir los datos y luego seleccione del menú, la opción deseada. Todos los procedimientos para análisis y gráficos despliegan ventanas, para que el usuario defina las características del análisis de los datos (variables a utilizar, estadísticas a calcular, niveles de significancia, etc) .



Cargando datos en la plantilla del Analyst

Además de poder digitar los datos en la plantilla de datos del Analyst, también se pueden cargar datos que ya están grabados como archivos de trabajo SAS.

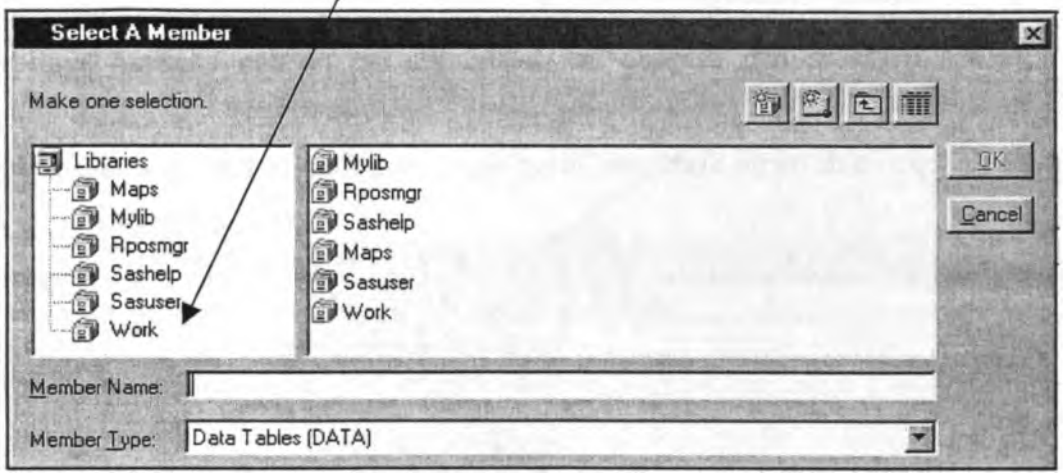
A continuación se presenta un ejemplo, paso a paso, en el cuál se realiza un análisis de regresión para un conjunto de datos con el Analyst. Los datos en la plantilla, se cargan de un archivo de trabajo SAS.

Paso 1: Crear el archivo de trabajo SAS. Se digita en el EDITOR, y se ejecuta el programa para crear el archivo de trabajo SAS. Este archivo es temporal, por lo tanto la librería donde se graba es WORK.

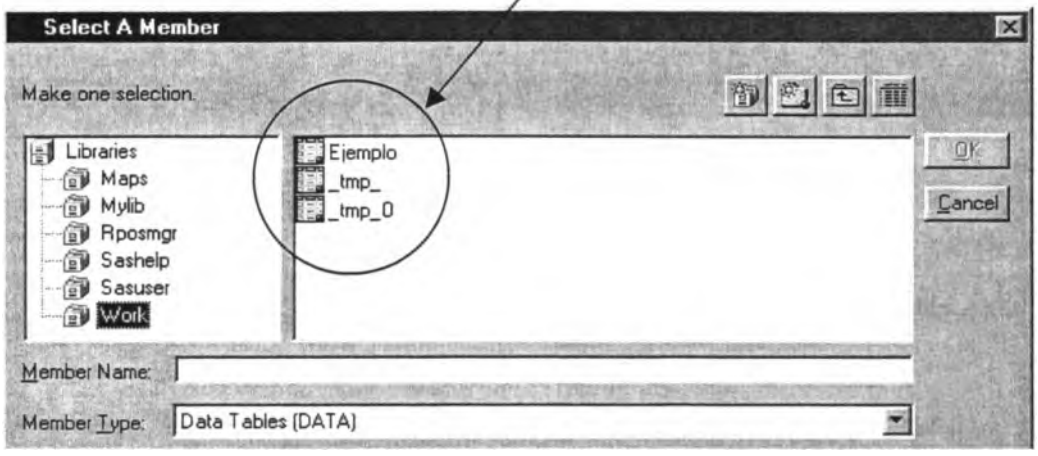
```
data Ejemplo;
input peso altura @@;
cards;
10.5 158.3 12.7 168.7  8.6 149.8 12.4 154.7 15.3 178.1
14.2 160.0  9.5 145.2 12.8 161.3 17.2 175.3 13.0 164.9
;
run;
```

Paso 2: Una vez creado el archivo de trabajo SAS, entrar al Analyst (como se explicó anteriormente), hasta obtener la ventana con la plantilla para entrada de datos. De click sobre la opción de menú **File** y después de click sobre la opción **Open By SAS Name**.

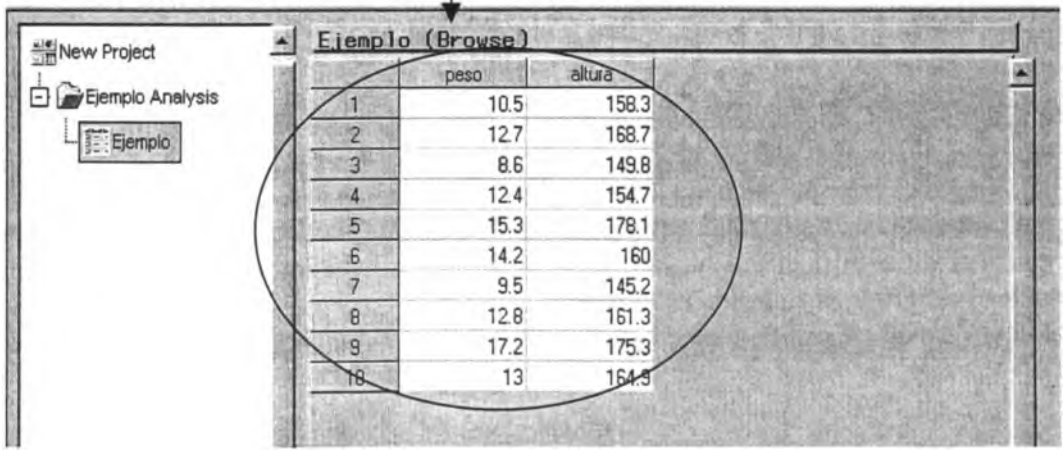
Paso 3: Seleccione de las librerías, la **Work**



Al dar click sobre Work, se muestra a la derecha de la ventana, los archivos de trabajo **SAS que están en** esta librería.



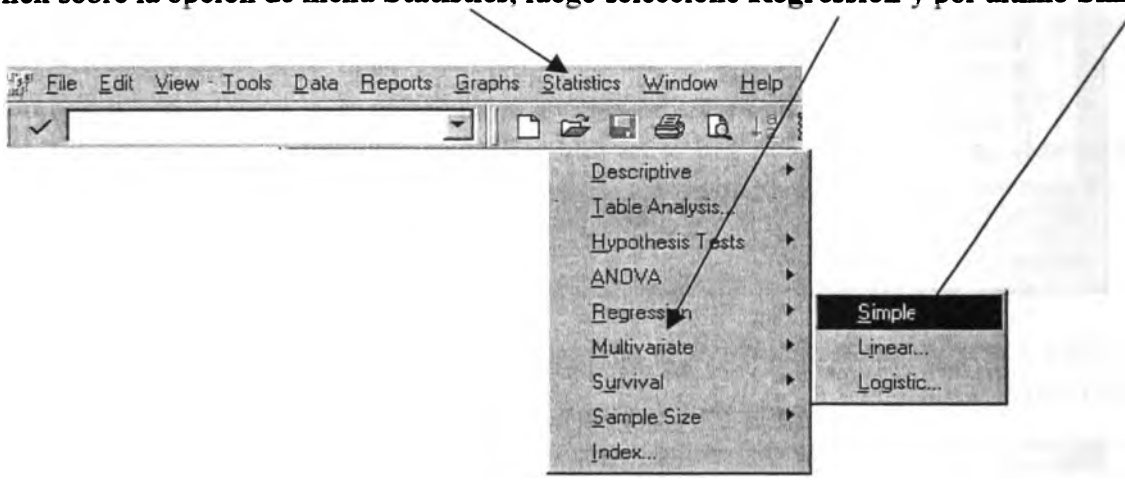
Paso 4: De doble click sobre el archivo deseado (Ejemplo en este caso), e inmediatamente los datos contenidos en este archivo, se cargan en la plantilla de datos del Analyst.



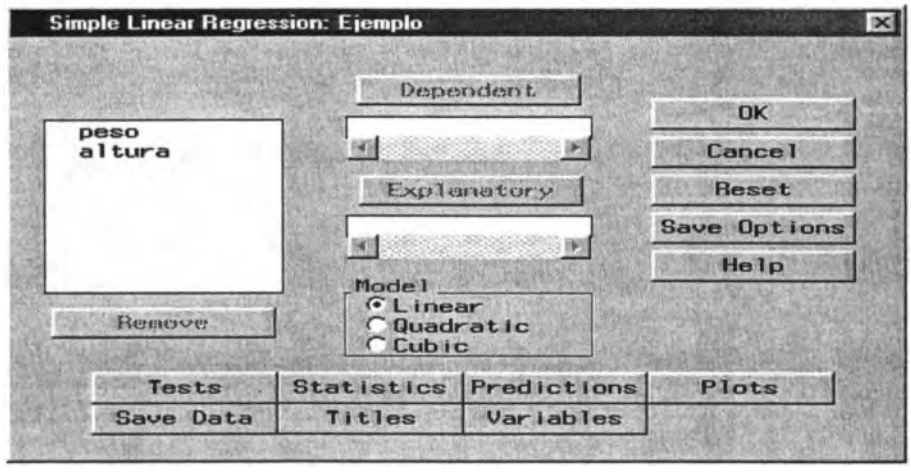
Procesando datos con el Analyst

Una vez que los datos se han cargado en la plantilla, se pueden analizar, graficar, etc. A continuación se presenta como calcular la regresión de Altura en función de Peso:

Dar click sobre la opción de menú **Statistics**, luego seleccione **Regression** y por último **Simple**:



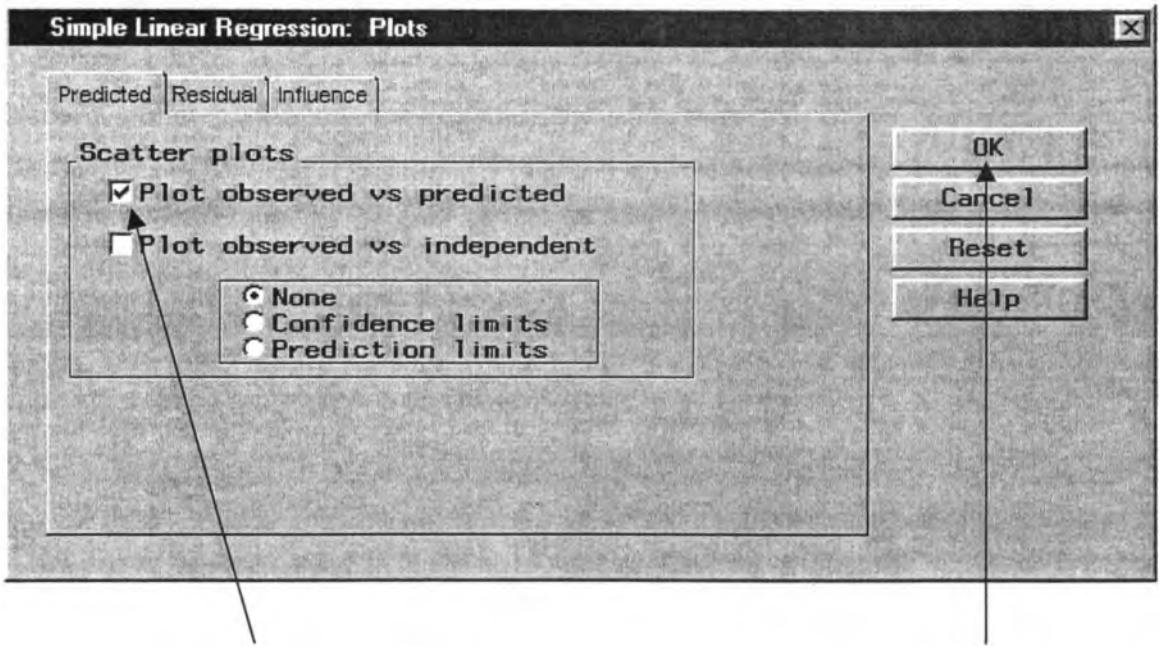
Aparece la siguiente ventana:



Seguidamente, indique la variable dependiente. Para esto, de click en **altura** y luego en **Dependent**. Para indicar la variable independiente, de click sobre **peso** y luego en **Explanatory**.

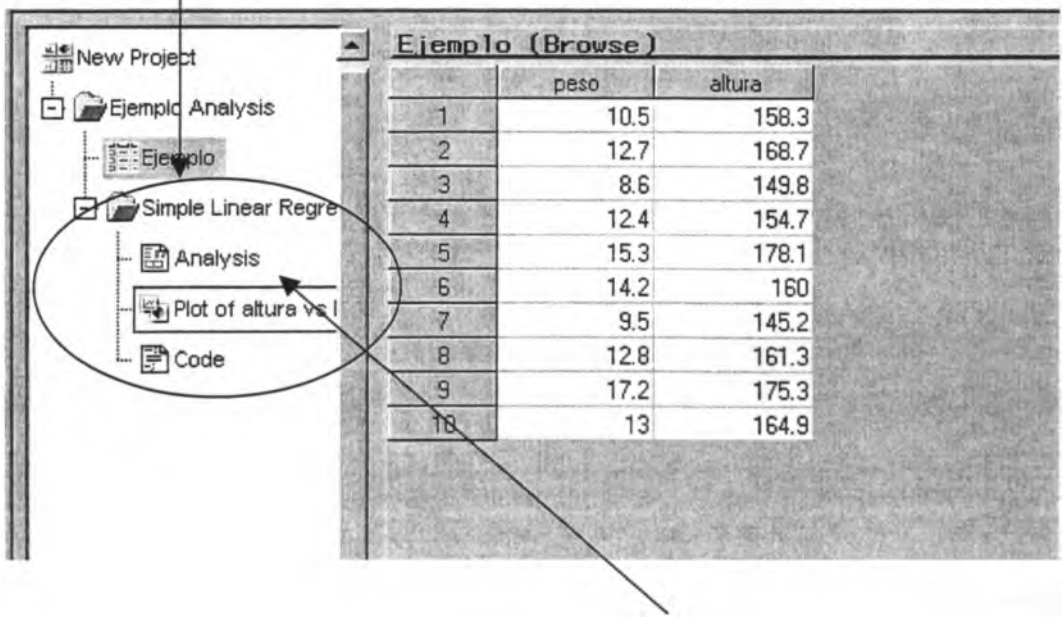
Esta ventana presenta una serie de opciones que al usuario le pueden interesar, como **Test**; donde se define el nivel de significancia, **Statistics**; donde se definen los parámetros a estimar, **Plots**; donde se indica que gráficos se quieren obtener.

Para este ejemplo, se calcula la regresión y se hace un gráfico de valores observados vs valores predichos. Se da click, sobre el botón Plots y se presenta la siguiente ventana:



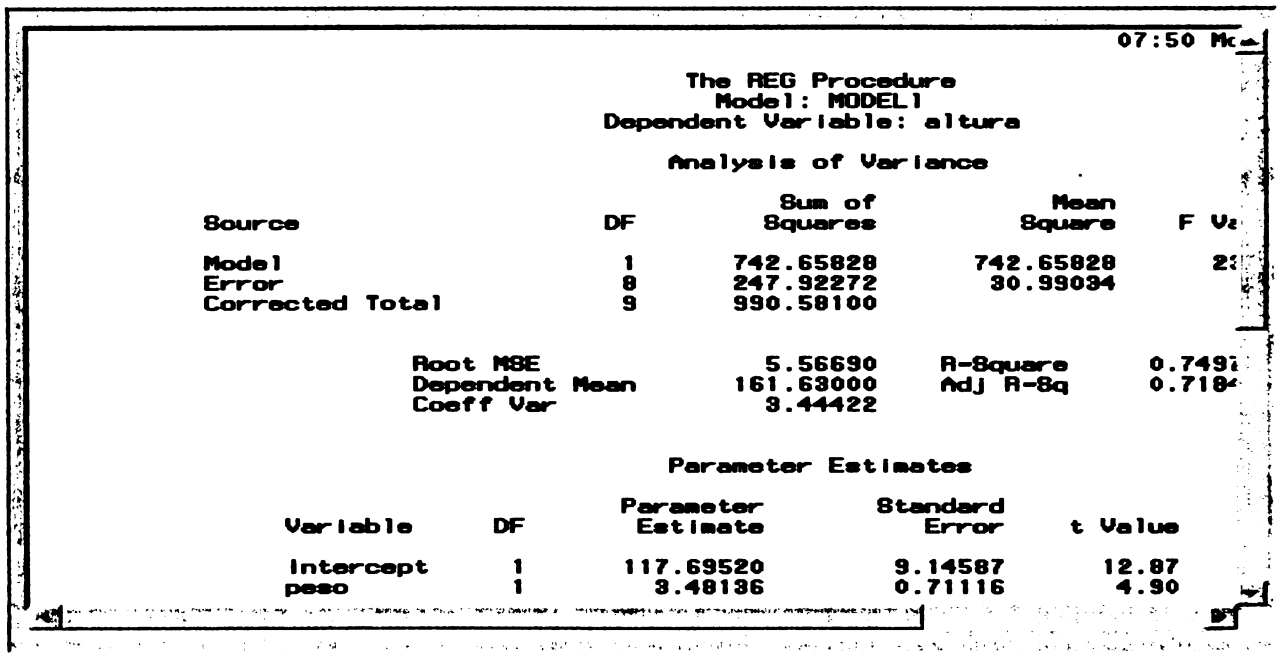
Se da click sobre **Plot observed vs predicted** y luego se da click sobre el botón **OK.**

Se regresa a la ventana anterior, y se da click sobre el botón **OK**. Seguidamente, el Analyst realiza el análisis de regresión y el gráfico solicitado. Se regresa al ambiente del proyecto, donde se muestran los análisis y gráficos que se realizaron.



Para ver la regresión o el gráfico, de click sobre **Analysis o Plot of altura vs peso**

Seguidamente se presentan los contenidos de las ventanas, al dar click sobre los componentes de análisis del proyecto. Si da click sobre **Analysis**, se muestra la siguiente ventana:



07:50 Mc

The REG Procedure
Model: MODEL1
Dependent Variable: altura

Analysis of Variance

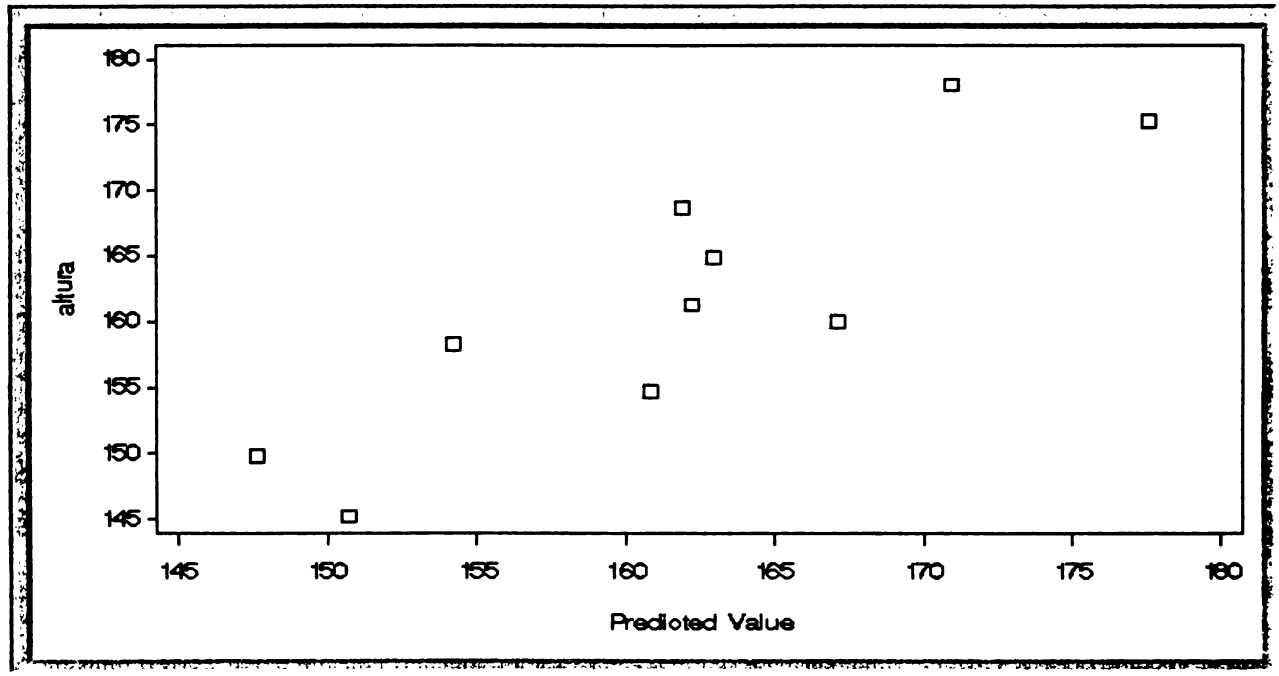
Source	DF	Sum of Squares	Mean Square	F Va
Model	1	742.65828	742.65828	23
Error	8	247.92272	30.99034	
Corrected Total	9	990.58100		

Root MSE	5.56690	R-Square	0.7493
Dependent Mean	161.63000	Adj R-Sq	0.7184
Coeff Var	3.44422		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value
Intercept	1	117.69520	9.14587	12.87
peso	1	3.48136	0.71116	4.90

Si da click sobre **Plot of altura vs peso**, se muestra la siguiente ventana



Puede imprimir las ventanas con los resultados, o bien, guardarlas en archivos. Para esto, utilice los íconos correspondientes o a través del menú .

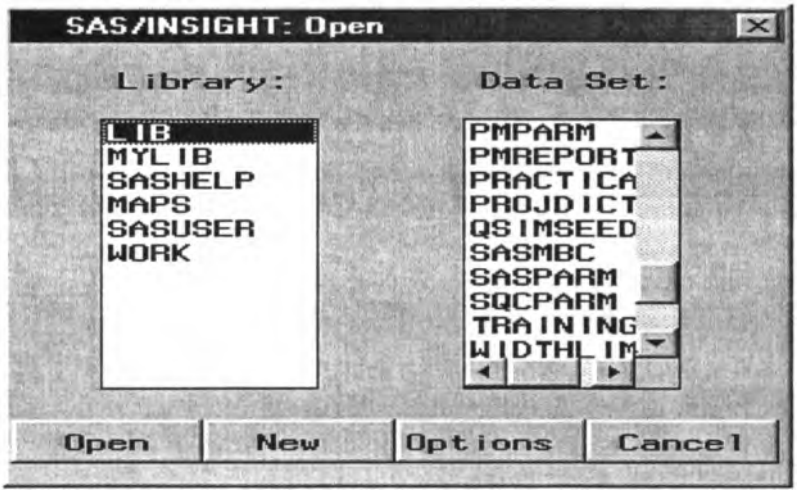
Análisis de datos interactivamente

El **Data Interactive Analysis**, es una herramienta que le permite a los usuarios realizar una serie de gráficos y algunas estadísticas de manera muy rápida y fácil. Para utilizar esta herramienta, los datos deben estar grabados en archivos de trabajo SAS.

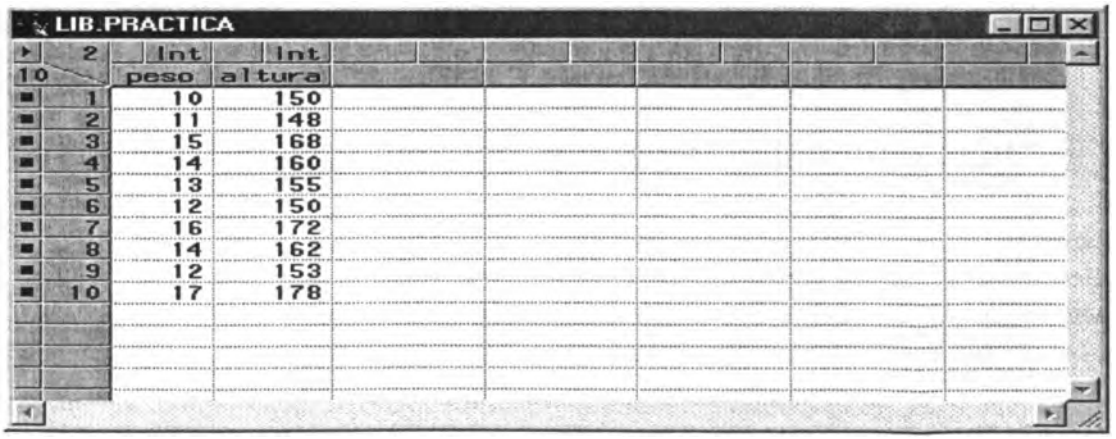
Para entrar al Data Interactive Analysis, siga los siguientes pasos:

- De click en la opción de menú **Solutions**
- En el siguiente menú, de click sobre la opción **Analysis**
- Seguidamente, de click sobre la opción **Data Interactive Analysis**

A continuación el SAS le presenta la siguiente ventana

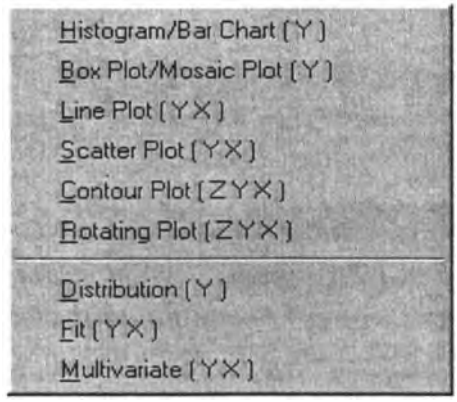


Esta ventana es para que el usuario indique la librería y el archivo que se va a trabajar. Si el archivo es temporal, la librería será WORK, si es permanente, la librería será la generada por el usuario. En este caso se va a seleccionar la librería **LIB** y el archivo **PRACTICA**. Esta librería y este archivo fueron creados anteriormente por el usuario. Seguidamente, de click sobre el botón **Open**. Se presenta la plantilla con el contenido del archivo

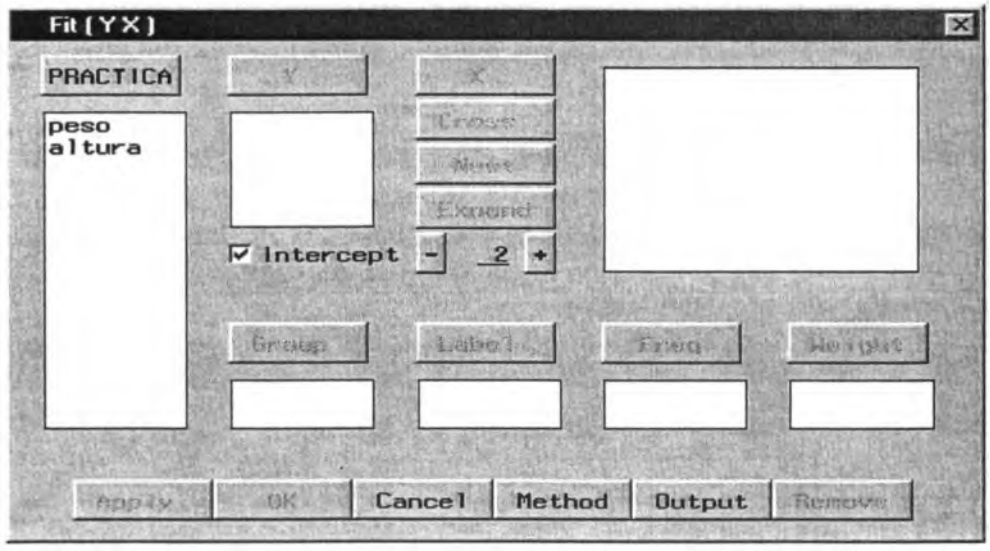


Se tiene disponible las opciones de menú **File, Edit** y **Analyze: Fill**

Si da click sobre Analyze, se despliega el menú con los diferentes gráficos y estadísticas que se pueden realizar:

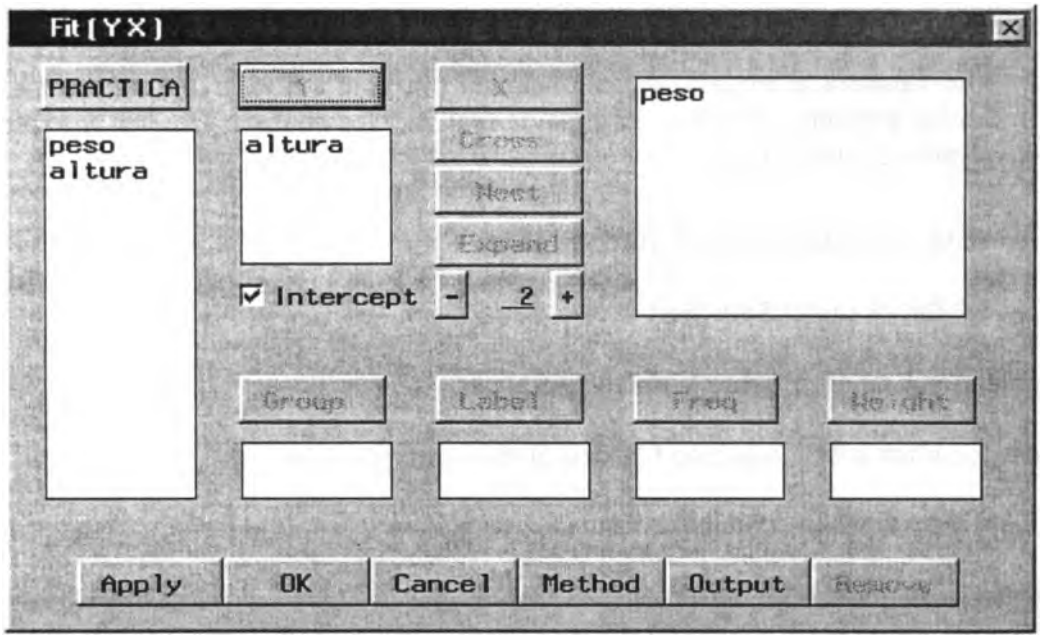


A manera de ejemplo, se va a calcular una regresión (**Fit (Y X)**) de la altura en función del peso. Al dar click sobre esta opción, se despliega la siguiente ventana:

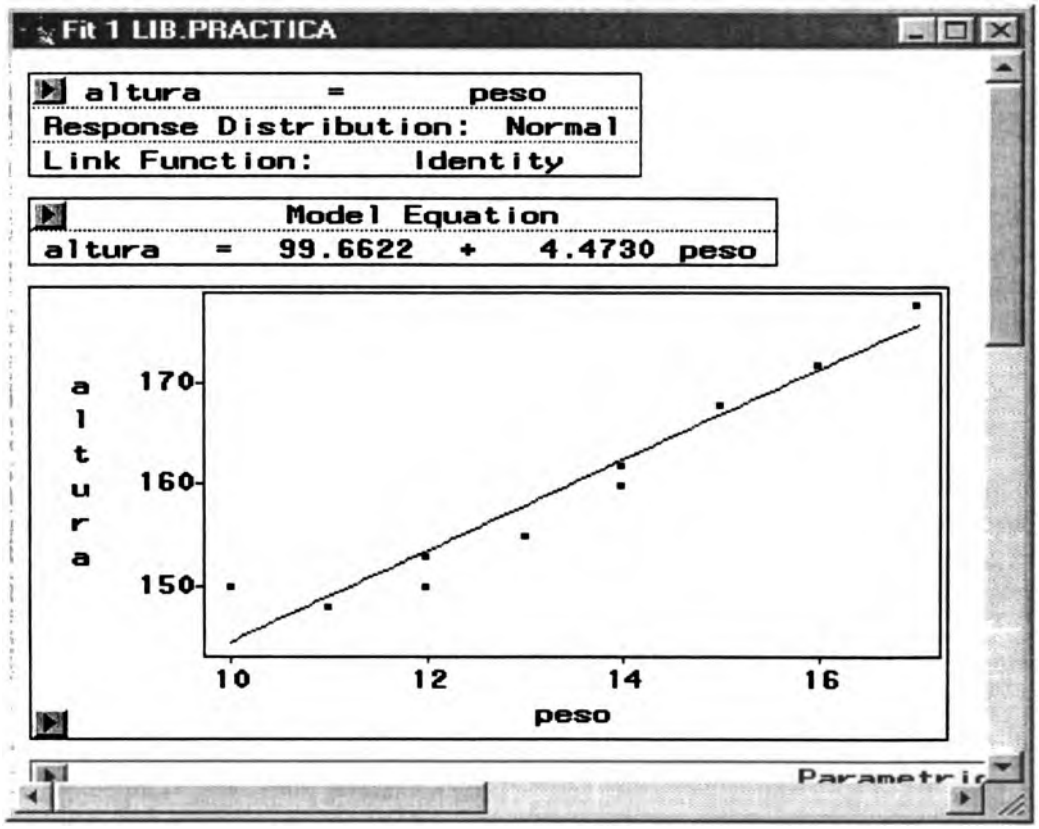


Seguidamente, se debe dar click sobre cualquiera de las variables. Como puede notar, los botones para seleccionar Y , X, así como algunas características sobre la regresión (Group, Label, etc) , no están accesibles. Al dar click sobre cualquier variable, estos botones se vuelven accesibles. Defina cual es la variable Y (dependiente) y cual es la variable X (independiente) y otras características.

A continuación la nueva situación de la ventana:



Para definir la variable Y, se dio click sobre altura y luego se presionó el botón **Y**. De igual manera se procedió con la variable **X**. Una vez definidas las variables, para obtener los resultados de la regresión, de click en **Apply** y el SAS le despliega la ventana con los resultados:



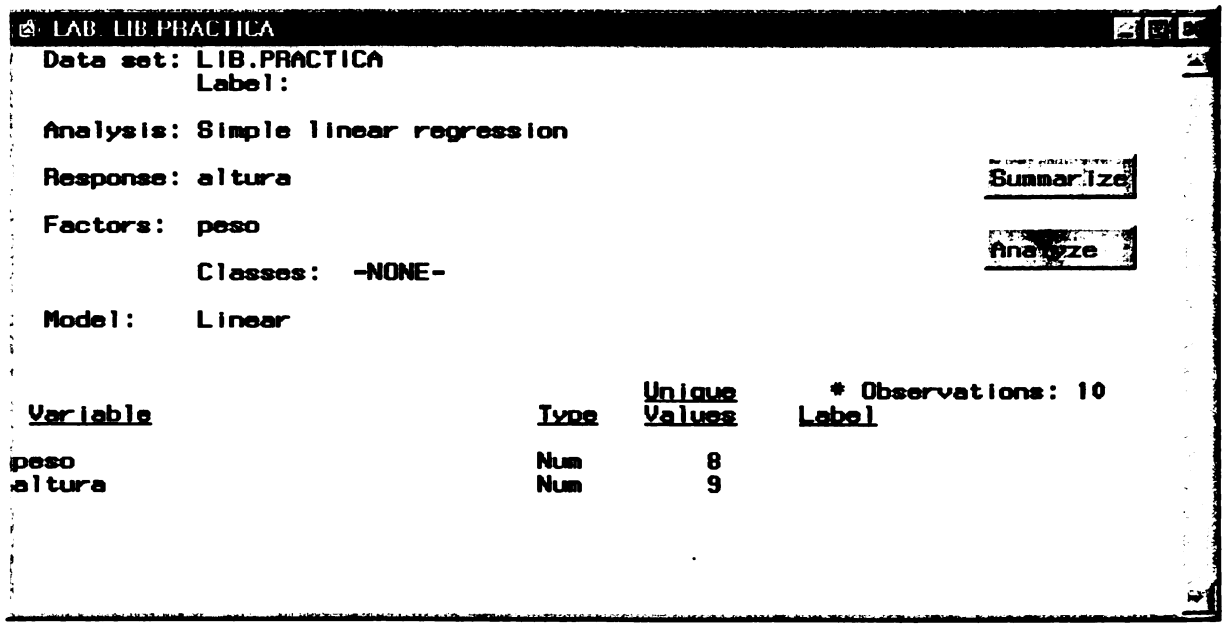
Análisis de datos guiados (Guided Data Analysis)

El **Guided Data Analysis**, es una herramienta que le permite a los usuarios calcular regresión lineal simple, análisis de varianza a una vía, regresión lineal múltiple e histograma con curva de probabilidad de normalidad para una variable. Además de realizar los análisis, también interpreta los resultados y da sugerencias a los usuarios.

Para entrar al Guided Data Analysis, siga los siguientes pasos:

- De click en la opción de menú **Solutions**
- En el siguiente menú, de click sobre la opción **Analysis**
- Seguidamente, de click sobre la opción **Guided Data Analysis**

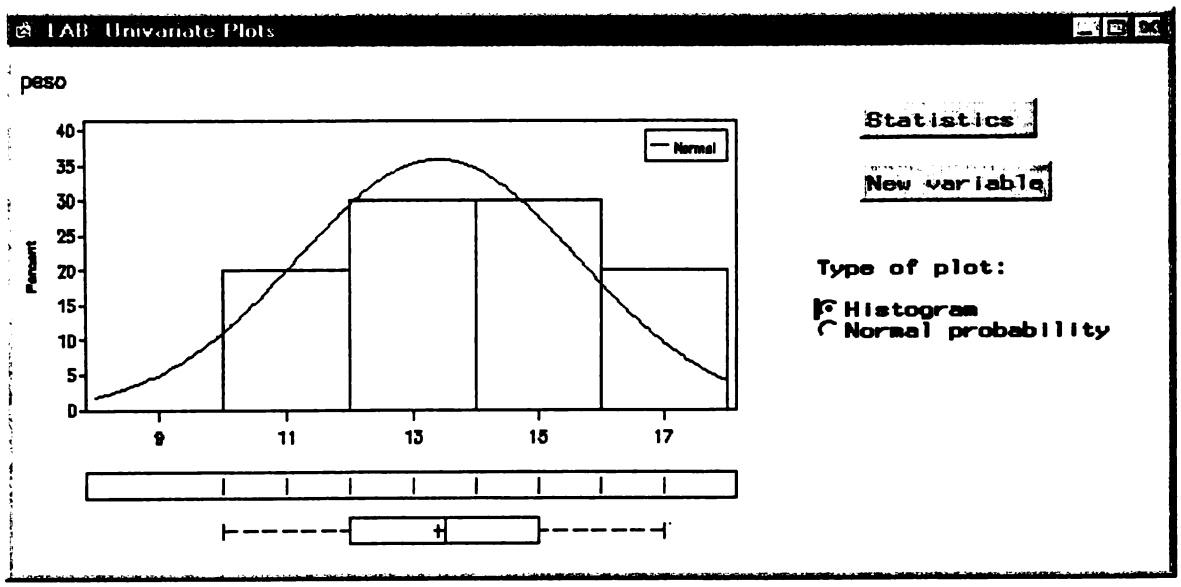
A continuación el SAS le presenta la siguiente ventana:



En esta ventana se indica el archivo que se va a trabajar (**Data set**), el tipo de análisis deseado (**Analysis**), la variable de respuesta (**Response**), la variable independiente (**Factors**) y el tipo de modelo de regresión (**Model**). Para realizar esto, simplemente se da click sobre el nombre de la opción.

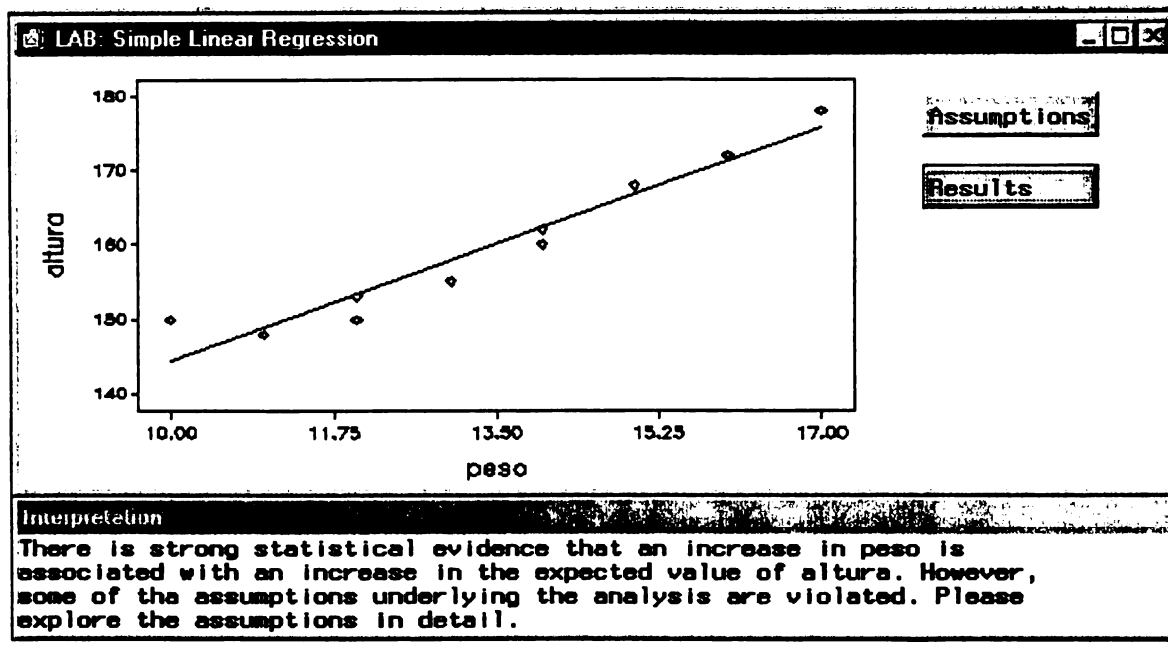
Para este ejemplo, se utiliza el archivo Practica, el cual reside en una librería permanente (LIB), y se hace un análisis de regresión simple.

Para obtener el histograma y/o la probabilidad de normalidad, de click sobre la variable en la parte inferior de la ventana. A continuación el resultado:



Si desea obtener el gráfico de probabilidad, de click sobre la opción **Normal probability**. Si quiere obtener las estadísticas descriptivas sobre la variable y la prueba de normalidad, de click sobre el botón **Statistics**.

Para obtener el análisis de regresión, de click sobre el botón **Analyze** de la ventana principal. Los resultados obtenidos se muestran a continuación:



Como puede observar, el Guided Data Analysis de una interpretación de los resultados. En este caso indica que hay evidencia estadística que el incremento del peso esta asociado con los valores de altura. Sin embargo, dice que algunos supuestos del análisis se están violando. Para ver los