GUIA DE APLICACION DE PRUEBAS ESTADISTICAS EN EL PROGRAMA SYSTAT 7.0 PARA CIENCIAS BIOLOGICAS Y FORESTALES

Resumen de un curso de estadística dictado por
Todd Fredericksen y José Carlos Herrera

Documento Técnico 86/1999

José Carlos Herrera F.

Laura E. Carse

Contrato USAID: 511-0621-C-00-3027-00
Chemonics International
USAID/Bolivia
Enero, 2000

Objetivo Estratégico de Medio Ambiente (USAID/Bolivia)

Guía de Aplicación de Pruebas Estadísticas en el Programa Systat 7.0 para Ciencias Biológicas y Forestales

Proyecto de Manejo Forestal Sostenible
BOLFOR

Cuarto Anillo
esquina Av. 2 de Agosto
Casilla 6204
Teléfonos: 480766 - 480767
Fax: 480854
e-mail: bolfor@bibosi.scz.entelnet.bo
Santa Cruz, Bolivia

BOLFOR es un proyecto financiado por USAID y el Gobierno de Bolivia e implementado por Chemonics International, con la asistencia técnica de Tropical Research and Development y Wildlife Conservation Society

TABLA DE CONTENIDO

SECCION I
INTRODUCCION

En esta guía se consideran pasos generales para ordenar y manipular datos, aplicar pruebas estadísticas e interpretar los resultados de investigaciones biológicas, ecológicas, y/o forestales. Estos procesos ayudarán en su trabajo cotidiano a estudiantes, profesores y asesores biólogos, forestales y agrónomos. Para ellos, se recomienda y se describe el uso de algunas herramientas o comandos de SYSTAT y EXCEL[1].

Para cada prueba estadística se describen: definición de la prueba, enfoque del problema, objetivo de la investigación, planteo de las hipótesis (nula y alternativa), forma de introducir los datos (en planillas electrónicas), diseño del muestreo (mas metodología), forma de ejecutar la prueba en SYSTAT, interpretación de los resultados, conclusión del ejemplo y notas importantes sobre las pruebas.

Esta Guía se basa en el curso de estadística que fue dictado por Todd Fredericksen y José Carlos Herrera en el Proyecto BOLFOR durante dos días, en febrero de 1999. En consecuencia, la mayoría de los conceptos, sugerencias, ejemplos y recomendaciones sugeridos y explicados por los disertantes, se menciona en esta guía.

[1] BOLFOR no pretende de ninguna manera hacer propaganda para estos paquetes.

SECCION II
ALGUNOS PASOS PRELIMINARES PARA APLICAR ESTADISTICA

Un investigador, para aplicar la estadística debe considerar los siguientes pasos:

1. Plantear una pregunta que se responderá durante el trabajo.
2. Plantear objetivos que responderán a las preguntas de interés. Estos, definen a las variables a analizar.
3. Plantear la hipótesis nula y alternativa con niveles de confianza de 0.05 ó 0.01. Estas permiten interpretar los resultados.
4. Realizar un diseño de muestreo aleatorio, sistemático y/o combinado. En este paso se definen las pruebas estadísticas que se aplicaran.
5. Recolectar datos en una planilla elaborada con sus respectivas leyendas que describan a las variables a estudiar.
6. Introducir los datos en planillas electrónicas. Las planillas deben ser elaboradas de acuerdo a los objetivos planteados y la planilla de campo.
7. Hacer una lista de las variables codificadas y no codificadas (tratamientos, bloques, lista de especies y sitios, etc.), para facilitar los análisis.
8. Corregir la base de datos utilizando algunas herramientas o comandos del SYSTAT (Sort, Select, etc.) y del EXCEL (ordenar, filtros automáticos, asistente para tablas dinámicas).
9. Describir los datos de manera gráfica y numérica (ESTADÍSTICA DESCRIPTIVA). Este paso es la última oportunidad para corregir los datos; ya que, trabajar con datos falsos puede llevar a conclusiones erróneas.
10. Inferir las características de la población basándose en las muestras (ESTADÍSTICA INFERENCIAL).
11. Interpretar los resultados de acuerdo a los objetivos e hipótesis planteadas.

Asumiendo que los pasos del 1 al 7 se han cumplido adecuadamente, a continuación se desarrolla en detalle el proceso desde la manipulación de datos hasta la interpretación de los resultados.

SECCION III
CORRECCION Y MANIPULEO DE LA BASE DE DATOS

Antes de utilizar las herramientas del "SYSTAT 7.0 for Windows" uno debe conocer las Ventanas y sus Menúes donde se ejecutan las pruebas descriptivas e inferenciales, que son las siguientes (Figura 1):

1. Main Windows, visualiza los resultados de las pruebas ejecutadas, como descripciones, análisis de los datos, etc.;
2. Data Windows, visualiza los datos en una planilla (en columnas y filas) donde pueden ser modificados (En SYSTAT un solo archivo se puede tener abierto al mismo tiempo).
3. Graph Windows, visualiza los gráficos automáticamente cuando se ejecuta en la ventana Main alguna prueba que requiera gráfico. Su edición puede realizarse en Main y Graph.
4. Command Editor, en esta ventana cualquier prueba se ejecuta a través de comandos (en esta guía no se utiliza esta ventana).

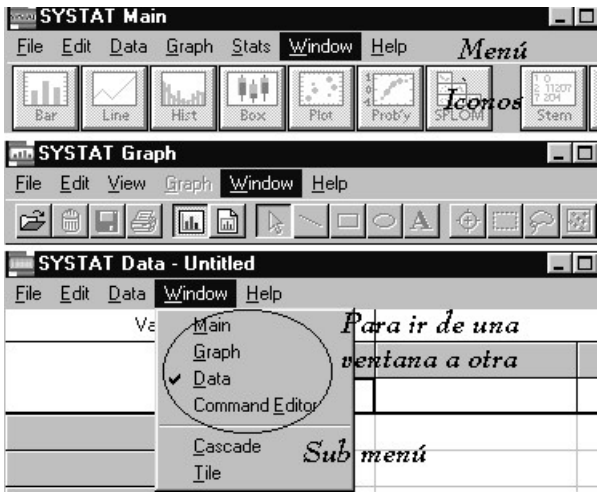


Figura 1

Las ventanas Main y Graph tienen menú e íconos y Data y Command solamente menú (Figura 1). El traslado de una ventana a otra se realiza a través del menú Window, que se encuentra en las cuatro ventanas; por ejemplo, cuando se ejecuta el menú Window, aparecen los nombres de las ventanas desplazados hacia abajo, y se debe señalar con el cursor en cual desea ejecutar (Figura 1). En algunos ejemplos se utilizarán las palabras ventana, menú, sub menú e íconos para ejecutar las pruebas estadísticas (Figura 1).

A. Base de Datos en el SYSTAT

La base de datos debe cumplir los siguientes requisitos:

Matriz: Cuando se introducen los datos en SYSTAT, se debe asegurar que la matriz de datos no tenga celdas vacías, porque este programa no las reconoce. Las celdas vacías deben llenarse con un punto (significa que no hay datos) o con un cero (significa que los datos tienen un valor).

Códigos y Números: Cuando a un dato se describe con dos palabras o una sola que tenga más de ocho caracteres, se deben codificar de manera legible y más corta, ejemplo: *Cebus apella* = Capella (cod), *Ateles paniscus* = Apanis (cod), Myrmecophagidae = Myrmeco (cod) (Figura 2), ya que, cuando se importa archivos en lenguaje ASCII, hacia el SYSTAT, los datos de una celda con dos palabras se convierten en dos columnas.

Los valores numéricos con decimales en SYSTAT deben ser asignados con puntos y no con comas porque si no el programa ejecuta como una variable calificativa (alfabética o como caracter).

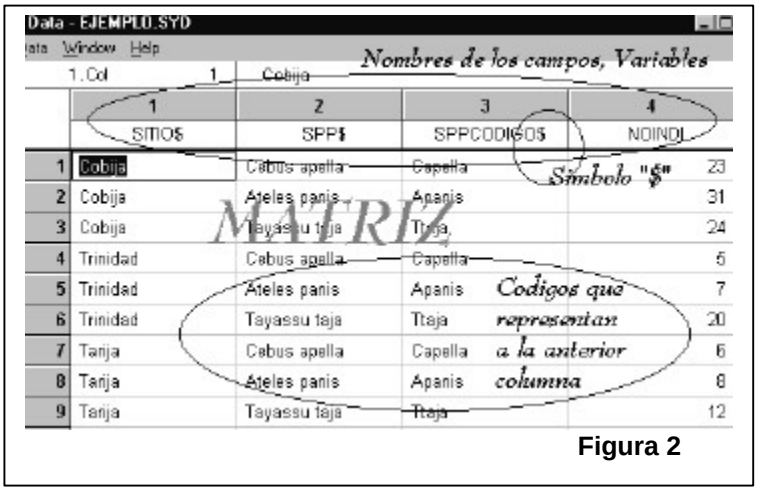


Figura 2

**Símbolo "$"**: En la palabra que encabeza la columna (nombre de campo), al final de ella, se coloca el signo "$"; Ejemplo: SPP$, SITIO$ y SPPCODIGO$ (Figura 2); estos campos incluyen datos no numéricos, que son palabras (caracteres); mientras por debajo del nombre de los campos sin signo, como de NOINDI, se escriben datos numéricos solamente. Por lo tanto el signo "$" es un distintivo entre una variable ordinal e interválica.

Cuando se importa un archivo de EXCEL hacia SYSTAT, en EXCEL no es necesario colocar el signo "$" porque SYSTAT lo coloca automáticamente. Pero para introducir datos no numéricos directamente en SYSTAT, se debe colocar el "$" al final del nombre de campo.

B. Corrección de Datos

En EXCEL se puede corregir y resumir los datos por medio las opciones de Ordenamiento Alfabético, Filtros y Tablas Dinámicas (recomendable por su gran efectividad); mientras que en SYSTAT se utilizan las de Sort (Ordenar) y Graph (gráficos). Es importante corregir los errores ortográficos en variables nominales y ordinales (caracteres) porque el programa considera diferente por ejemplo: "desc." no es igual que "desc,"; y en caso de los datos números mal escritos, podrían influir en la distribución de la muestra por ejemplo: en una distribución de los diámetros de una especies, como 40, 0, 455, 45, 50, 47 y 43, no podría existir un especie con un diámetro de "455 ó 0" cm (influyen en la dispersión y medias).

B1. Ordenar en EXCEL

Esta función ordena en forma ascendente o descendente los datos numéricos y caracteres (palabras) situados en una columna o fila (Figura 3).

Ejecutar: ÷ **Datos** ÷ **Ordenar**

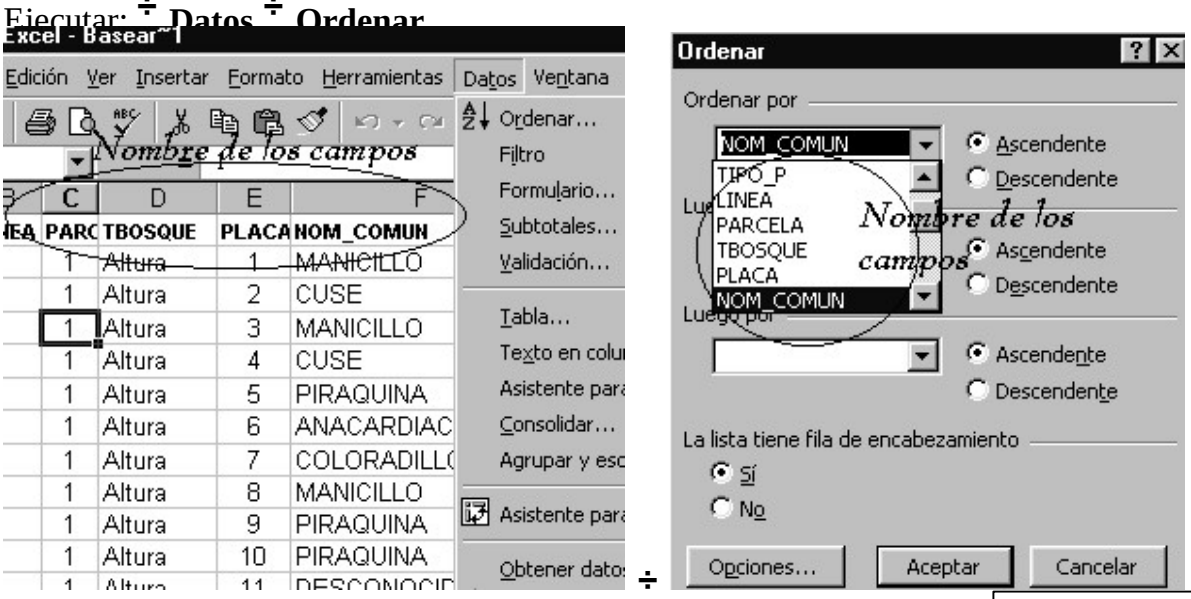


Figura 3

B2. Filtros Automáticos en EXCEL

Esta función filtra una variable o dato deseado de la base de datos con todos sus datos correspondientes. Cuando se ejecuta la función aparece una flecha al lado derecho del nombre de campo con una dirección hacia abajo; de ésta, se elige una variable colocando el cursor encima de la flecha y aparecen desplazados todas las variables que contiene la columna (Figura 4).

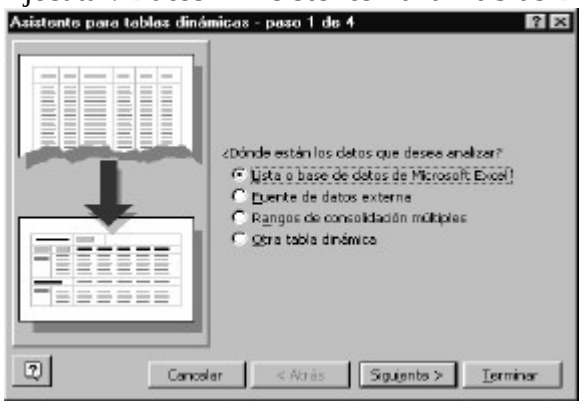
Ejecutar: **Datos** ÷ **Filtro** ÷ **Autofiltro**

Figura 4

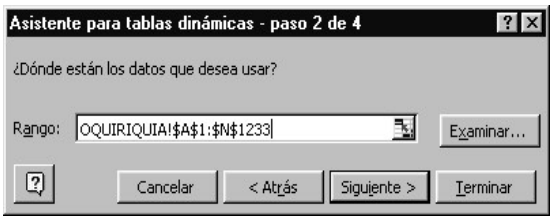
B3. Tablas Dinámicas (EXCEL)

Esta opción permite resumir y analizar la base de datos. Para utilizar las Tablas Dinámicas se debe recurrir al Asistente para tablas dinámicas, donde se especifica la base de datos ya sea externa o interna (paso 1), rango de la base de datos (paso 2), definición de los resultados (paso 3) y finalmente se sitúa los resultados (paso 4) ver figuras:

Ejecutar: **Datos ÷ Asistente Para Tablas Dinámicas**



1er Paso



2do Paso

3er Paso

4to Paso

Figura 5

En el primer paso se debe indicar si los datos se encuentran en EXCEL o fuera de este programa; en el segundo paso, se debe indicar qué rango ocupa la base de datos (entre columnas y filas); en el tercer paso, se debe indicar con qué variables se trabajará (nombres de los campos; de acuerdo a los Objetivos); y en el cuarto paso, se debe indicar dónde se desea colocar los resultados (recomendable colocar en la misma hoja pero pasando dos columnas vacías, después de la base de datos).

En una base de datos es recomendable trabajar con una sola tabla dinámica. Para modificar los resultados generados por una tabla dinámica ya no es necesario iniciar el proceso desde el primer paso, si no sólo desde el tercer paso donde se puede modificar y finalizar; para esto es necesario dejar el cursor dentro la tabla dinámica y ejecutar: **Datos ÷Asistente Para Tablas Dinámicas**. Además, en cualquier variable que se encuentra dentro de FILA, COLUMNA y PAGINA (paso 3) se pueden omitir las palabras que están dentro de una columna, mientras con las variables numéricas en DATOS (paso 3) se pueden realizar operaciones como

sacar suma, conteo, promedio, desviación estándar, máximo, mínimo, porcentaje, etc. Para esta operación, se debe colocar el cursor encima, en cualquiera de las mencionadas opciones y hacer doble clic con el botón derecho del ratón y luego aparecerán las opciones mencionadas.

B4. Ordenar (SYSTAT)

El comando Sort (ordenar) hace que los datos numéricos o alfabéticos se ordenen de manera ascendente o descendente en la columna (con esta operación se obtiene el mismo resultado que con EXCEL), ver figura 6.

Ejecutar: **Datos ÷ Sort**

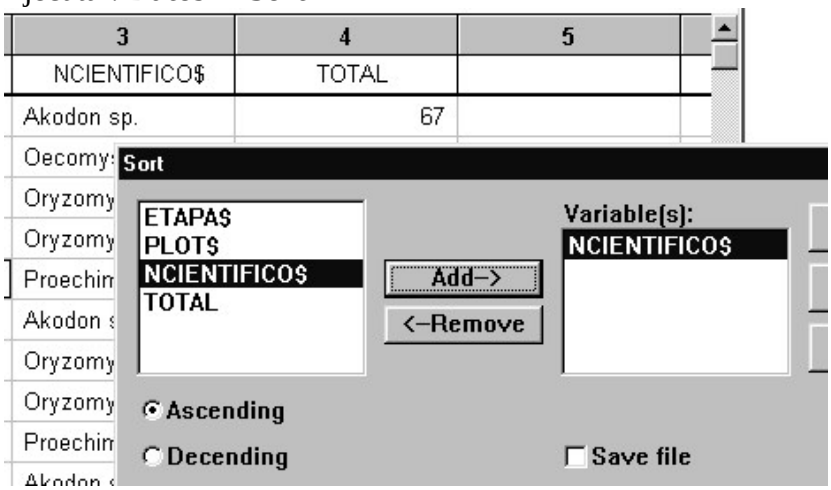


Figura 6

Ejecutar: **File ÷ Open ÷ Data**.

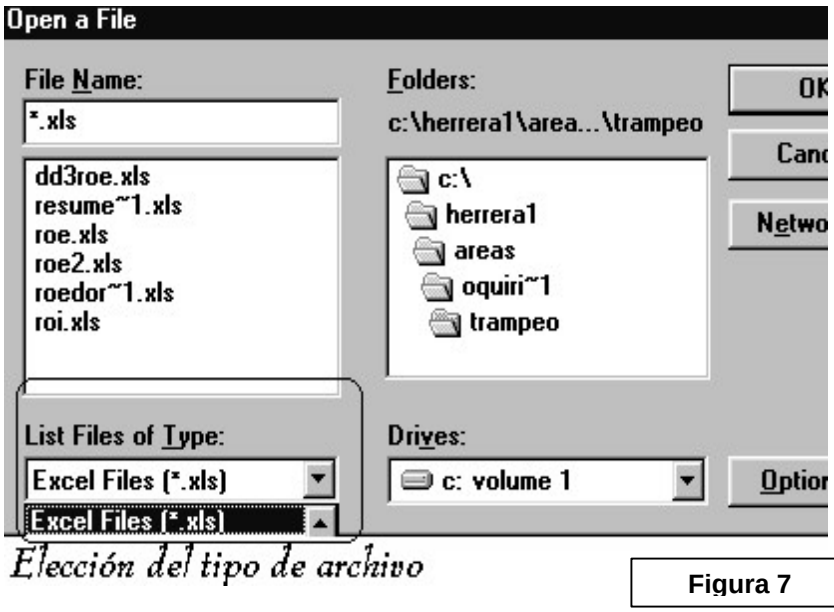


Figura 7

C. Importar Datos

SYSTAT puede importar archivos de EXCEL y de Quattro PRO con su propia extensión pero de una versión antigua de ellas. Sin embargo, el formato mas recomendable para esta operación es texto (*.txt) por su uso universal (ASCII).

Para importar se deben considerar las indicaciones del párrafo de códigos y números (Figura 7).

SECCION IV
LINEAMIENTOS PARA LA SELECCION DE PRUEBAS ESTADISTICAS

La estadística es una herramienta que ayuda a describir a una muestra (ESTADISTICA DESCRIPTIVA) y a través de ésta, se infieren los resultados a una población general (ESTADISTICA INFERENCIAL). En general se debe conocer la base de datos, posteriormente describir la muestra y recién hacer inferencias (Figura 4). Para inferir se usan pruebas paramétricas y no paramétricas y antes de usar, cada una de las pruebas, se deben cumplir ciertas características como se indica en el Cuadro IV-1 y en Figura 8.

Cuadro IV-1: Requisitos que deben cumplir las pruebas paramétricas y no paramétricas.

Características	Pruebas Paramétricas	Pruebas No Paramétricas
Distribución	Normal (simétrico) y variancias homogéneas	Libre (asimétrico y varianzas heterogéneas o desiguales)
Observaciones	Reales	Reales o convertidos a rangos
Variables	Intervalicas o proporcionales	Nominales, ordinales, escala de intervalo
Centralización	Promedio (media)	Medianas, modas
Dispersión	Varianzas	Rangos
Conteos	Son apropiados para datos que muestran frecuencias	Deben ser transformados
Tamaño de la muestra	N > 30	N < 30

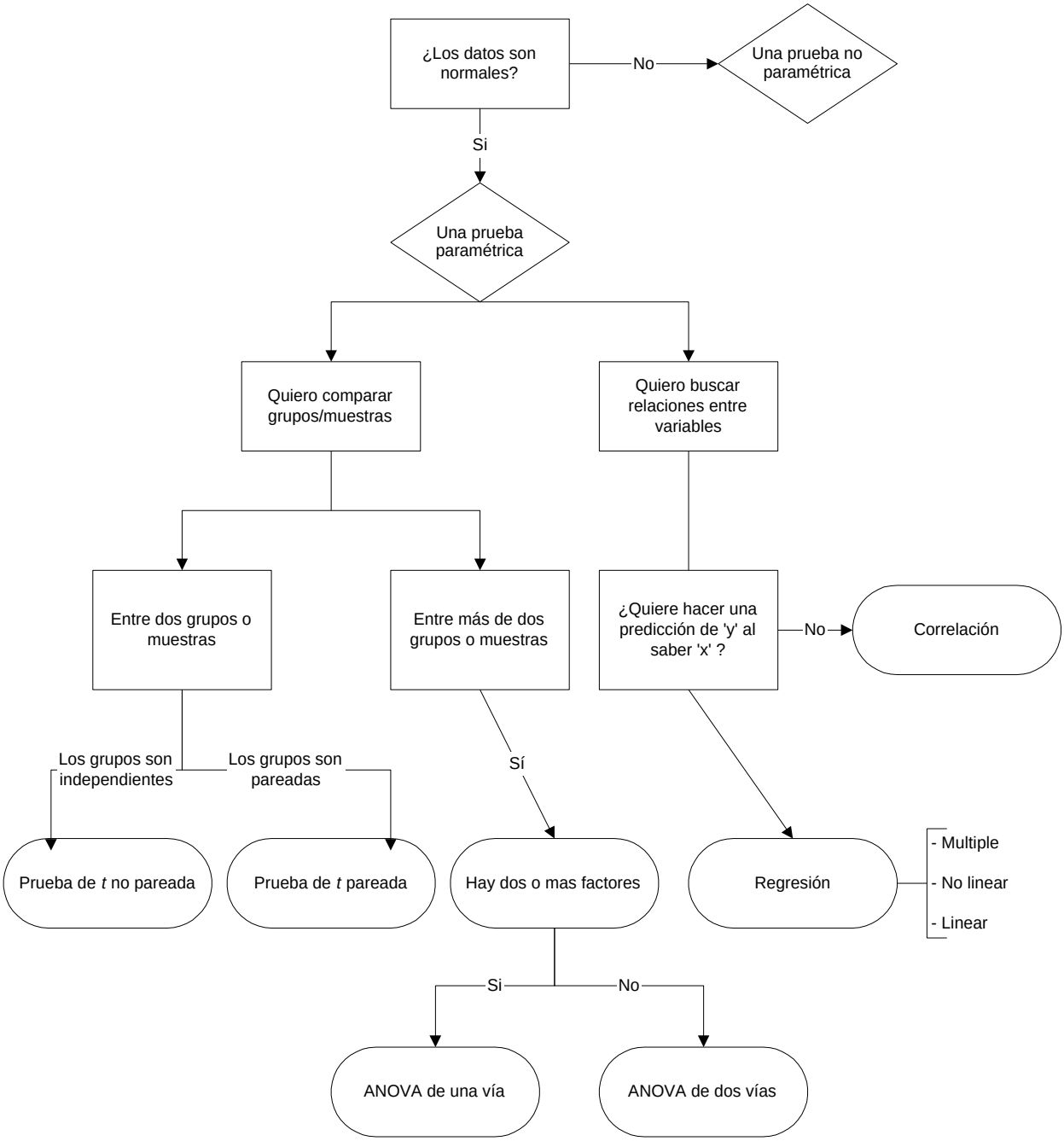


Figura 8. Flujo de pruebas paramétricas. En caso de que no cumplan los requisitos para aplicar una prueba paramétrica, se debe recurrir a una segunda opción, hacer transformaciones para que cumplan los requisitos, y si aun no cumplen los requisitos, el análisis se realiza con pruebas no paramétricas.

A. Estadística Descriptiva

En EXCEL y SYSTAT se realizan pruebas estadísticas descriptivas e inferenciales, pero en el primero, sólo se pueden realizar pruebas sencillas y básicas, mientras, en el segundo, se pueden realizar pruebas con más detalle, y las más avanzadas, con facilidad. En esta guía sólo se describen algunas opciones de EXCEL , y de SYSTAT se detallan más pruebas y con mayor claridad.

Los parámetros para describir una muestra son: de tendencia central (media, mediana, moda) y de dispersión (desviación estándar, rango). Estos parámetros se pueden representar por medio de gráficos, y con este fin los más utilizadas son: el de caja, tallo y hoja, normal, histograma, barras y curva acumulada.

A1. Gráficos log – normal en SYSTAT

Ejecutar: **Graph ÷ Density Function**

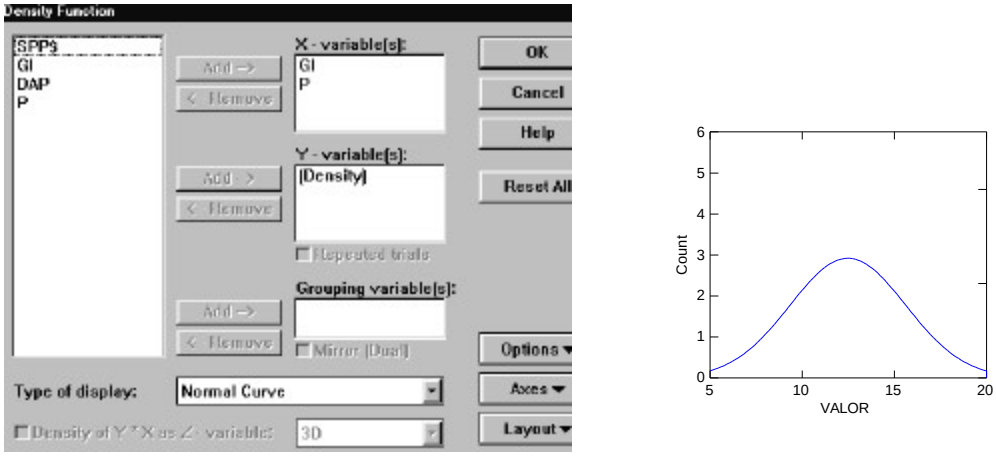


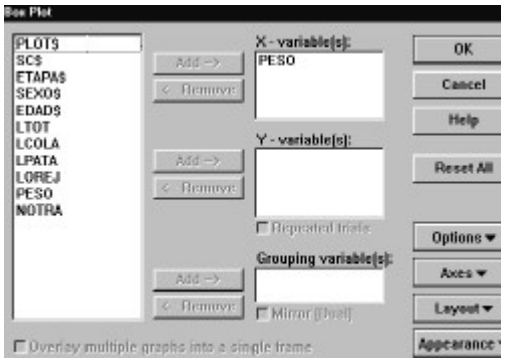
Figura 9

Una distribución se comporta de acuerdo a los valores de los datos, y generalmente, existen tres tipos: distribución normal, cuando los valores de los datos extremos son menores y en menor cantidad que en el medio; distribución con cola negativa, cuando existen más datos mayores que menores (mayor cantidad de datos con ceros); y distribución con cola positiva, la muestra se comporta de manera contraria a la anterior (Figura 9). A las dos últimas generalmente, se las denomina Log - normal. Estas distribuciones ayudan a determinar qué prueba se aplicará (PARAMETRICA o NO PARAMETRICA).

A2. Gráfico de Caja

El gráfico de caja es otra forma de presentar la distribución de datos. La ubicación de la media en la caja indica el rango de los datos. Cuando este gráfico no tiene una raya o línea al medio de la caja (mediana) y/o bigotes en uno o en los dos extremos, quiere decir que la distribución no es normal (Figura 10).

Ejecutar: **Graph ÷ Box Plot**



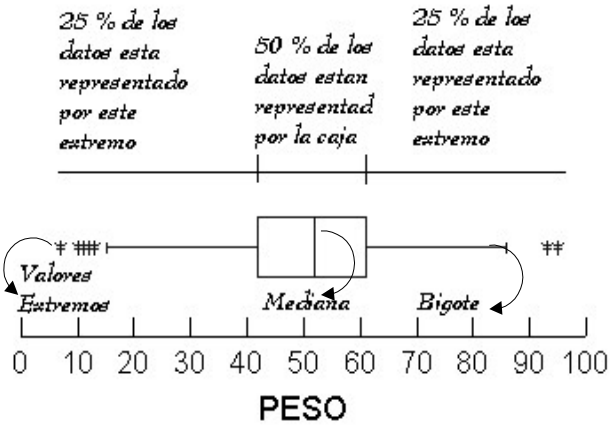
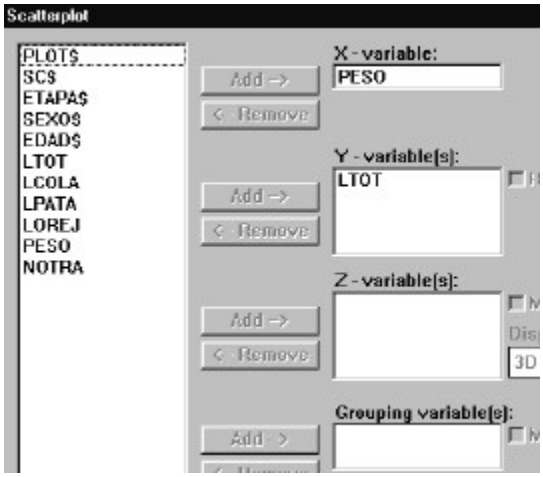


Figura 10

A3. Diagrama de Puntos

Este gráfico visualiza la relación entre dos variables. Se usa para realizar una prueba de correlación y regresión. También son importantes para presentar resultados de las pruebas mencionadas (Figura 11).

Ejecutar: **Graph ÷ Scatterplot**



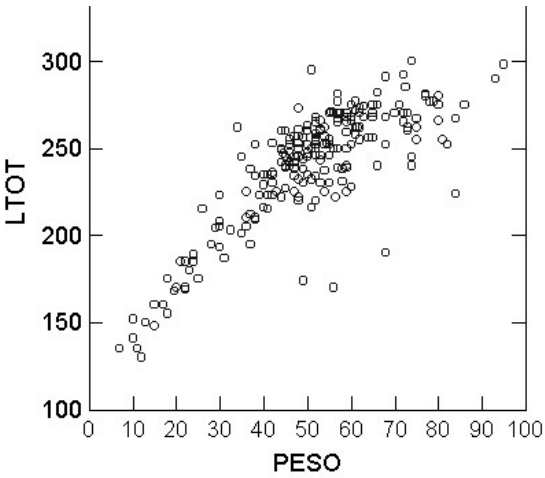
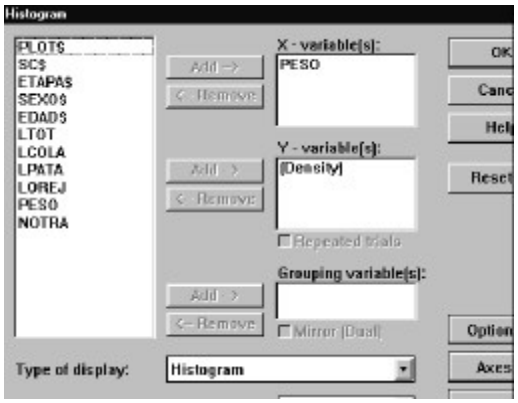


Figura 11

A4. Histograma

Este gráfico muestra si la distribución de los datos, es normal o no (Figura 12).

Ejecutar: **Graph ÷ Histogramar**



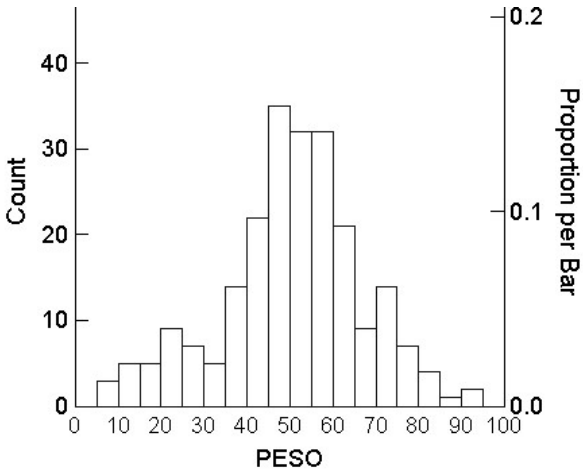


Figura 12

A5. Diagrama de Tallo y Hoja

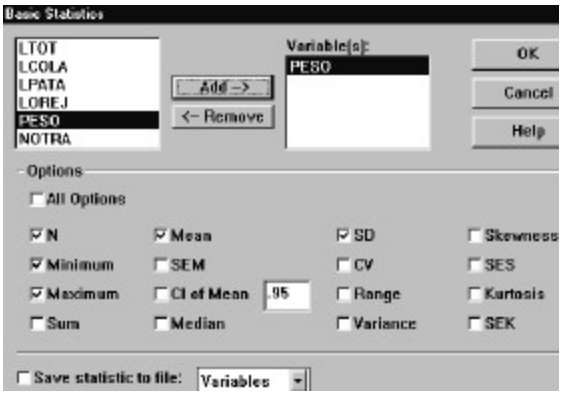
En este gráfico uno puede identificar los valores máximos, mínimos y extremos. También muestra todos los valores (tal como son) y a través de ellos uno puede detectar cuáles datos se encuentran fuera de la distribución ya que son marcados con “Outside values” (Figura 13).

Ejecutar: **Stats ÷ Descriptive Stats ÷ Stem ans Leaf**

Figura 13

A6. Estadística Básica o Descriptiva

Ejecutar: **Stats ÷ Descriptive Stats ÷ Basis Stats**



	PESO
N of cases	227
Minimum	7.000
Maximum	95.000
Range	88.000
Sum	11577.500
Median	52.000
Mean	51.002
Standard Dev	16.664
Variance	277.698
C.V.	0.327

Figura 14

Por medio de la opción "Descriptive Stats" se describen los parámetros, que representan la distribución de una muestra en forma numérica.

En la ventana que se visualiza existen 16 opciones para sacar estadísticas descriptivas o básicas. Se pueden elegir a cualquiera o a todas las opciones (All Options) señalando encima de las opciones, con el cursor del ratón (Figura 14).

B. Estadística Inferencial

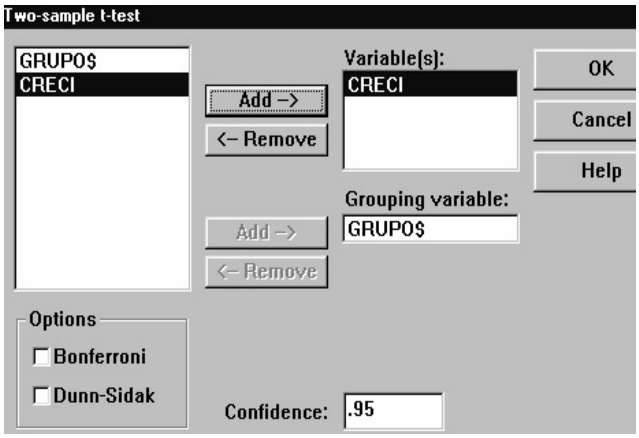
B1. Pruebas Paramétricas

La forma de colocar la base de datos para cada prueba se muestra en ANEXO 1.

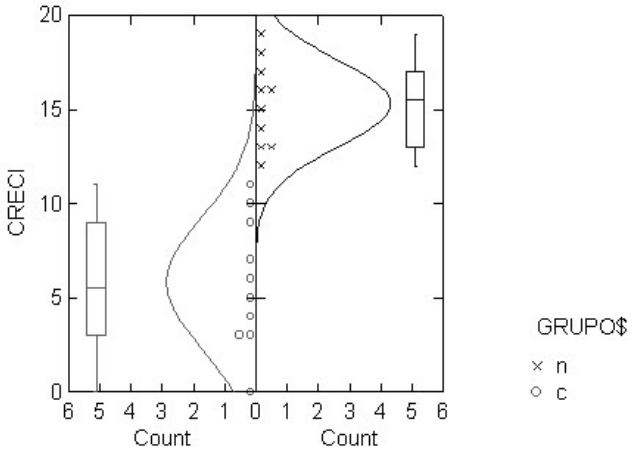
B1a. *t* Para dos Grupos Independientes

Definición: se usa para comparar los promedios de una variable entre dos poblaciones diferentes. **Objetivo**: determinar el crecimiento de las plántulas con y sin abono. **Hipótesis**: H_0 = no hay una diferencia entre el crecimiento de una población con tratamiento y sin tratamiento; H_a = existe diferencia entre los dos tratamientos. **Diseño**: 10 plantas se trataron con abono de nitrógeno (n) y 10 plantas sin abono (c). El crecimiento fue medido después de una semana. **Interpretación:** En el gráfico (Figura 15), de ambos grupos *c* y *n*, la distribución normal y de caja no se hallan paralelos horizontalmente; lo que significa que hay relativa diferencia entre ambas distribuciones de las muestras (caso contrario existiría relativa similaridad). En los resultados numéricos se muestra los valores de los tratamientos "*c*" y "*n*" (*Group*) y son: número de las muestras (*N*), media (*Mean*) y desviación estándar (*SD*), y estos son los parámetros que caracterizan a la muestra.

Ejecutar: **Stats ÷ t-test ÷ Two groups**



Resultados/gráficos



Resultados/numéricos

```
TEST CRECI * GRUPO$ / BONF
Two-sample t test on CRECI grouped by GRUPO$
  Group                N          Mean            SD
    c                 10         5.800         3.490
    n                 10        15.300         2.312
Separate Variance t = -7.177  df = 15.6    Prob = 0.000
                      Bonferroni Adjusted Prob = 0.000
Difference in Means = -9.500   95.00% CI = -12.312 to -6.688

Pooled Variance   t = -7.177  df = 18      Prob = 0.000
                      Bonferroni Adjusted Prob = 0.000
Difference in Means = -9.500   95.00% CI = -12.281 to -6.719
```

Figura 15

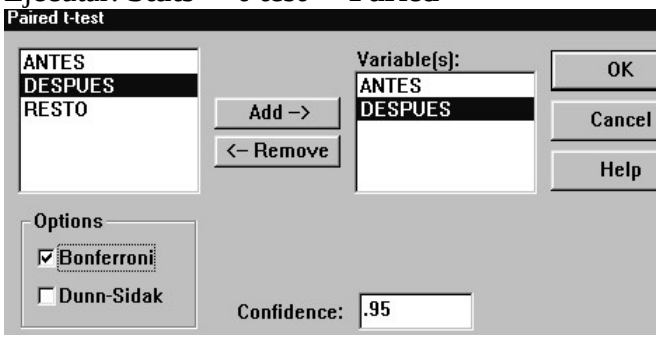
Por debajo de estos párametros o matriz se encuentran dos opciones: *Separate variance* t *test* y *Pooled variance* t *test*. El primero se usa cuando las varianzas de las muestras no son iguales; y el segundo cuando las varianzas son similares. En ambos se encuentran el valor de *t*, grados de libertad (*df*), probabilidad (*Prob*), diferencia de medias (*Difference in Means*) e intervalos de confianza (*CI*), este último indica que la media se encuentra entre -12.281 y –6.719, con una certeza de 95%.

Conclusiones: Los árboles del grupo '*c*' tienen un crecimiento de 5.8 cm como promedio sin nitrógeno y el grupo "*n*" tratados con nitrógeno crecieron 15.3 cm. Estas diferencias de promedios indican que las distribuciones de las poblaciones inferidas son diferentes, *P=0.000* (=P<0.0001). Con estos resultados se apoya la hipótesis alternativa (Figura 15).

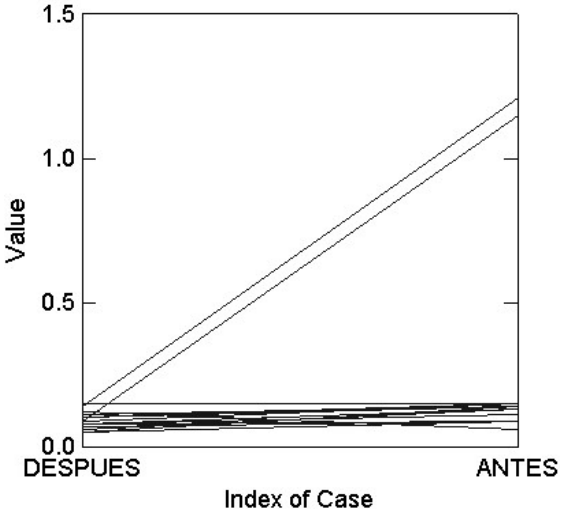
Nota:

- Cuando la muestra es menor a seis (n<6), por más que tengan una distribución normal, los resultados son poco fiables (podría existir sesgo), sin embargo, muchos libros mencionan que cuando n < 30 no tiene una distribución normal.
- Cuando la probabilidad es igual a cero (resultados en SYSTAT), no significa un valor absoluto, sino que la probabilidad es menor a 0.001.

Ejecutar: **Stats ÷ t-test ÷ Paried**



Resultados/gráficos



Resultados/numéricos

```
>TEST ANTES DESPUES / BONF

Paired samples t test on ANTES vs DESPUES with 15 cases
  Mean ANTES        =  0.262
  Mean DESPUES      =  0.091
    Mean Difference =  0.171   95.00% CI = -0.031 to 0.373
      SD Difference =  0.365           t =  1.812
                                      df = 14      Prob = 0.091
```

Figura 16

B1b. Prueba de *t* Pareada

Definición: Se usa para comparar los promedios de dos muestras pareadas. La prueba se emplea en diseños previos y posteriores (antes y después), sobre los mismos individuos o unidades muestrales.

Objetivo: Determinar si las lianas interfieren en la presión de agua que sube hacia la copa de árboles individuales.

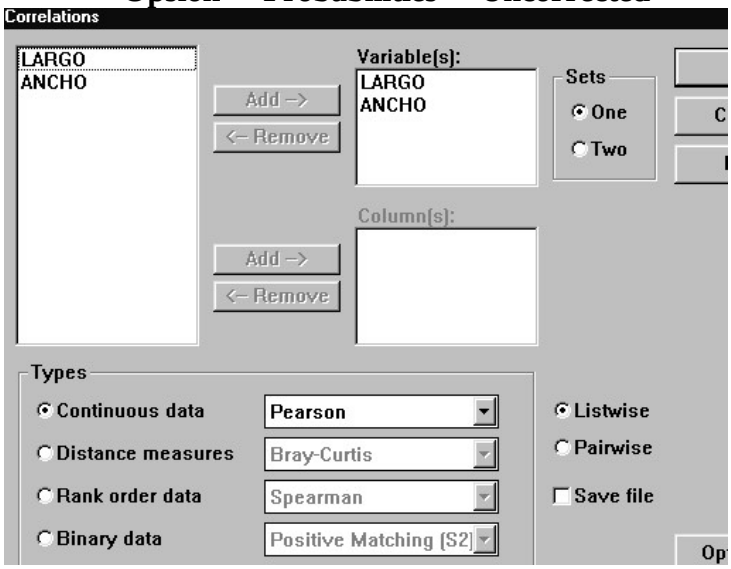
Hipótesis: H_o = no hay una diferencia de la presión de agua de los árboles antes y después de cortar lianas; H_a = contraria a la anterior.

Diseño: A 15 árboles se le midió la presión de agua, luego se cortaron las lianas y después de una hora se volvió a medir la presión de agua.

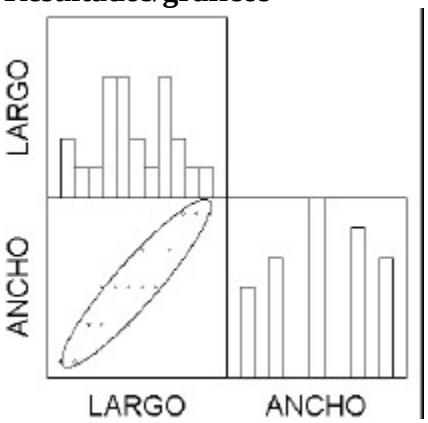
Interpretación: En los resultados gráficos se muestra que dos de los árboles tenían mayor presión antes de la corta de lianas (de dos líneas su extremo de la derecha es más elevado). En los resultados numéricos, el valor de *Mean antes* es mayor que *Mean después*; *Mean Difference* indica la diferencia de los promedios entre los dos grupos; *95% CI* indica que la media se encuentra entre los intervalos -0.031 y 0.373 (con un error de 5%

de que el promedio que quede fuera de este rango); *SD Difference* indica la diferencia de desviación estándar de ambas mediciones; *t* indica un valor que se puede encontrar en la tabla de *t*-Student para ver la probabilidad; *df* indica los grados de libertad con que se ingresa a una tabla de t-student; *Prob* (P) = indica la probabilidad de que la prueba sea similar o diferente y en este caso hay similaridad entre las dos poblaciones. **Conclusiones**: Existe la posibilidad de que la presión sea igual entre antes y después de la corta de lianas. Lo que significa que las lianas no afectan a la presión de agua que sube por los troncos hacia la copa de los árboles (Figura 16).

Ejecutar: **Stats ÷ Correlations ÷ Simple**
Opsion ÷ Probabilities ÷ Uncorrected



Resultados/gráficos



Resultados/numéricos

```
>CORR >PEARSON LARGO ANCHO / PROB

Pearson correlation matrix
                    LARGO        ANCHO
 LARGO              1.000
 ANCHO              0.925        1.000
Bartlett Chi-square statistic:    39.616 df=1 Prob= 0.000

Matrix of Probabilities
                    LARGO        ANCHO
 LARGO              0.0
 ANCHO              0.000        0.0
Number of observations: 23
```

Figura 17

B1c. Correlación

Definición: mide la relación de dos o más variables; ejemplo, relación entre el N° de hojas y la altura de la planta.

Objetivo: Determinar si hay una asociación entre el ancho y el largo del ala de una especie.

Hipótesis: H_0 = no hay una asociación entre largo y ancho del ala de las aves; H_a = hay una correlación positiva.

Diseño: Entre una especie y otra no hay una diferencia morfológica y sólo se puede diferenciar correlacionando el largo y ancho de la ala; donde: sp1, no tiene una correlación; y sp2, tiene una correlación positiva. Se midieron el largo y ancho del ala de 23 aves; entonces, ¿los individuos medidos a qué especie pertenecen?.

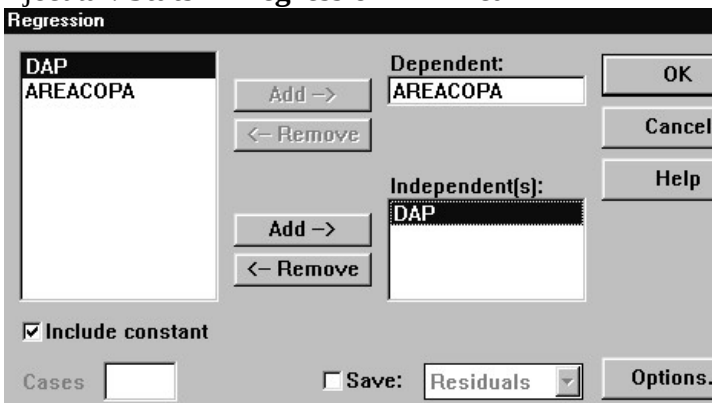
Interpretación: en los resultados gráficos se muestra una elipse inclinada hacia la derecha (formada entre ancho y largo) que indica que existe una relación positiva. En los resultados numéricos hay una matriz de correlación (*Pearson correlation matrix*) donde el valor de *r* de Pearson es igual a 0.925 lo que indica que hay una asociación fuerte (para valores cerca a ± 1, la asociación es fuerte, mientras cerca a cero la asociación es débil, en correlación lineal), y por debajo de la matriz se

encuentran *Bartlett Chi-square Statistic* que es la herramienta estadística usada para probar la hipótesis, grados de libertad (*df*) y la probabilidad Prob = 0.000 =P<0.0001; por debajo de la matriz de Pearson se encuentra la matriz de la probabilidad (*Matrix of Probabilities*) que indica la probabilidad o la certeza de la asociación, y también se encuentra el número de observaciones (*Number of observations = 23*).
Conclusión: Entre el ANCHO y el LARGO del ala existe una asociación positiva (r=0.925, P<0.001), por esta correlación se rechaza la hipótesis nula y se acepta la alternativa (Figura 17).

Notas:

- Se puede hacer una correlación entre muchas variables, pero cuando es mayor a diez se utiliza la correlación Bonferroni.
- Existen terminologías para el valor de r; ejemplo, cuando r se encuentra entre: 0 y 0.19 = la correlación es muy débil; 0.20 – 0.39 = la correlación es débil; 0.40 – 0.69 = hay una correlación moderada; 0.70 – 0.89 = hay una fuerte correlación; y 0.90 – 1.00 = la correlación es muy fuerte.

Ejecutar: **Stats ÷ Regression ÷ Linear**



Resultados/gráficos
Ejecutar: **Graph ÷ Scatterplot**

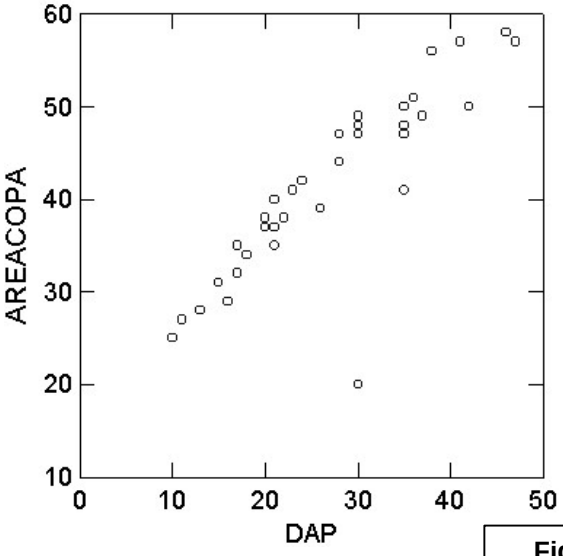
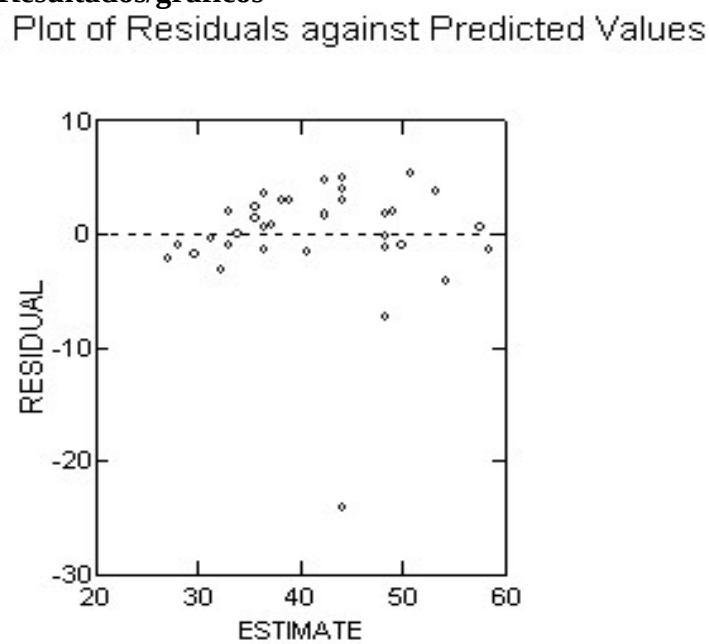


Figura 18

B1d. Regresión Lineal

Definición: Por medio de una ecuación una variable independiente (eje X = DAD) predice el comportamiento de una variable dependiente (eje X = AREACOPA) ver Figura 18. **Objetivo**: Determinar la ecuación para predecir el área de copa por medio de la medición del DAP de un árbol.
Hipótesis: H_0 = No hay una relación entre DAP y el área de copa; H_{a1} = Hay una relación significativa y marcada entre DAP y el área de copa. H_{a2} = Hay una relación significativa, pero no es muy marcada.
Diseño: Se midieron el DAP y el área de copa de 37 árboles.
Interpretación: En el gráfico *Scatterplot* se observa la asociación de los puntos que fueron diagramados por medio de DAP (independiente) y AREACOPA (dependiente); en este se muestran los puntos agrupados en forma de una línea inclinada hacia la derecha (correlación positiva).

Resultados/gráficos



Resultados/numéricos

```
>MODEL AREACOPA = CONSTANT+DAP; >ESTIMATE

Dep Var: AREACOPA  N:37  Multiple R:0.860 Squared multiple R:0.740
Adjusted squared multiple R:0.733 Standard error of estimate:4.939

Effect  Coefficient  Std Error  Std Coef Tolerance   t   P(2 Tail)
CONSTANT   18.754     2.392     0.0        .       7.840   0.000
DAP         0.842     0.084     0.860     1.000    9.983   0.000

                      Analysis of Variance
Source     Sum-of-Squares df  Mean-Square  F-ratio       P
Regression 2430.620       1     2430.620   99.656       0.000
Residual    853.650      35       24.390
          *** WARNING ***
Case 26 is an outlier        (Studentized Residual = -8.837)
Durbin-Watson D Statistic     1.523
First Order Autocorrelation   0.234
```

Figura 19

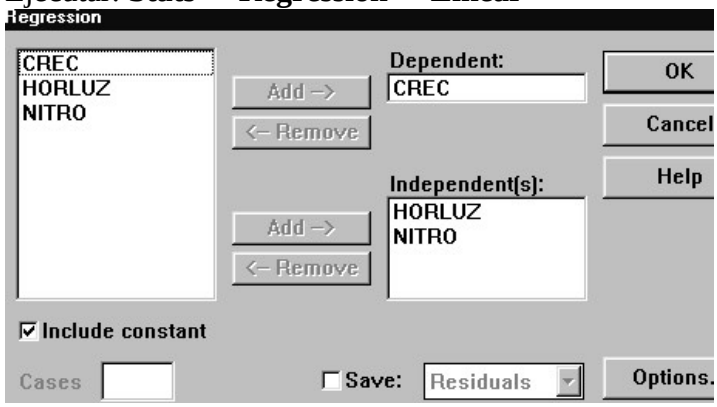
En los resultados gráficos, se observa los puntos residuales (que es la distancia entre el punto y la línea cero, y es calculada por el programa), estos puntos deben ser dispersos, caso contrario significa que la asociación es sesgada. En los resultados numéricos, en las primeras líneas se observa: número de muestras (*N*=37), R múltiple (*R*=0.86; se usa cuando hay una sola variable independiente asociada a una dependiente), cuadrado de R múltiple (Squared Multiple R=0.740, es la proporción de la varianza total que corresponde a la variable dependiente, además, explica la variabilidad de la dependiente), *Adjusted Squared Multiple* (*R* =0.733, se utiliza cuando se realizan análisis con más de una va-riable independiente), *Stan-dard Error of Estimate* (= 4.939, es una medida resi-dual del análisis de varianza y se usa con modelos que tienen más de una variable independiente). En la primera matriz se muestra el coeficiente, error estándar, coeficiente estándar, tolerancia (relevante cuando se analiza a más de una variable independiente) valor de *t* y el valor de *P* (probabilidad con 2 colas) que corresponden a CONSTANT y DAP. De la segunda matriz las más importantes son *F-ratio* y *P*; la primera se utiliza para probar la hipótesis, y la segunda indica la probabilidad de que sea correlacionada o no, entre variables.

Conclusiones: El valor de la correlación simple, entre las dos variables, está cerca a 1 (*Multiple R* = 0.860) y 74% de la varianza del área de copa es explicada por el DAP (*Squared Multiple R).* Entonces, para ver la relación se considera los valores de *r* y *P*. En este caso, según estos valores, podemos desechar la hipótesis nula y aceptar la segunda hipótesis. En este ejemplo existe una relación entre el área de copa y el DAP (F=99.656, P < 0.001) (Figura 19).

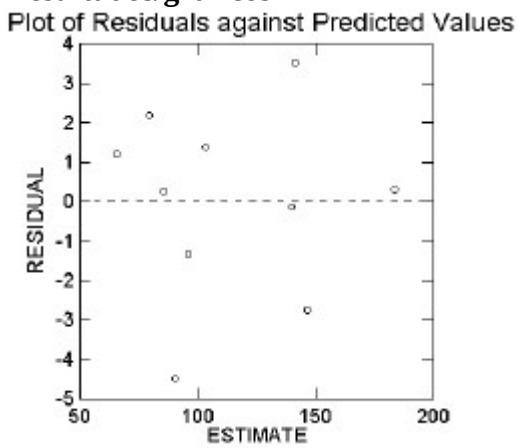
Nota:

- La parte proporcional de la varianza inducida por la variable independiente, se expresa mediante el coeficiente de la correlación y las pruebas de significación se obtienen a partir de la varianza total.
- La variable '*x*' es la que se debe controlar o la que se quiere pronosticar.
- Se puede expresar el valor de r^2 como un porcentaje (0.78 = 78%, explica la variabilidad entre *Y* y *X*).

Ejecutar: **Stats ÷ Regression ÷ Linear**



Resultados/gráficos



Resultados/numéricos

```
>MODEL CREC = CONSTANT+HORLUZ+NITRO >ESTIMATE

Dep Var: CREC  N:10  Multiple R:0.998  Squared multiple R:0.996
Adjusted squared multiple R:0.995  Standard error of estimate:2.678

Effect   Coefficient Std Error Std Coef Tolerance  t   P(2 Tail)
CONSTANT -139.925    11.116     0.0       .     -12.588   0.000
HORLUZ      0.333     0.010     0.797    0.951   32.690   0.000
NITRO       0.124     0.007     0.449    0.951   18.429   0.000

                            Analysis of Variance
Source     Sum-of-Squares   df  Mean-Square   F-ratio      P
Regression    12626.210     2     6313.105    880.493     0.000
Residual         50.190     7        7.170

*** WARNING ***
Case       5 is an outlier       (Studentized Residual = -2.348)

Durbin-Watson D Statistic     3.209
First Order Autocorrelation  -0.606
```

Figura 20

- Para asegurarse de que la regresión es consistente en la predicción de la variable, uno puede fijarse la variabilidad de las mediciones arriba y abajo de la línea de ajuste (residuales) debieron ser iguales, en otras palabras en la figura de residuales debería verse una nube de puntos sin ninguna tendencia lineal, curvilínea o agrupaciones.
- Antes de realizar un análisis de regresión, es bueno ver en un gráfico la distribución de la variable independiente y dependiente.

B1e. Regresión Múltiple

Definición: Con más de dos variables independientes se predice el compartimiento de una variable dependiente; a esta prueba, en ocasiones, se la entiende como una regresión no lineal. Objetivo: Pronosticar el crecimiento de una planta según su exposición a horas luz y la cantidad de nitrógeno que se encuentra en el suelo (Figura 20).

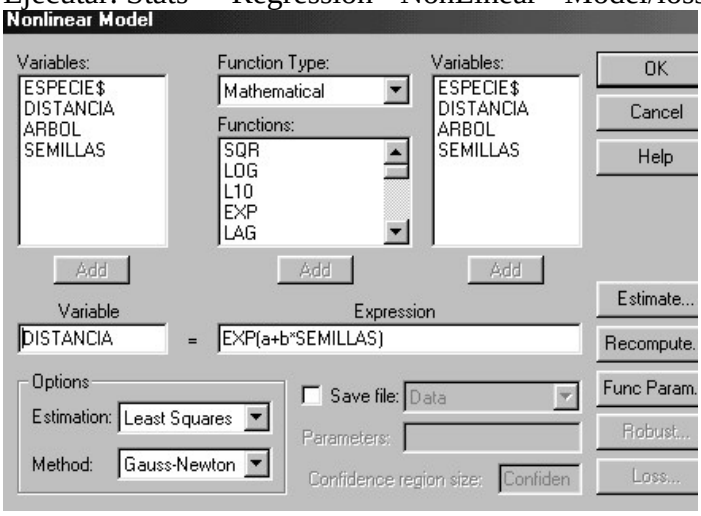
Hipótesis: H_0 = no hay ninguna relación entre el crecimiento y horas - luz y cantidad de nitrógeno; Ha = hay una relación entre ellas.

Diseño: A diez plantas se midieron horas de luz de exposición y la cantidad de nitrógeno existente en el suelo.

Interpretación: En los resultados gráficos se muestra el diagrama de los puntos residuales con respecto al punto cero (buena dispersión, no hay sesgo). En resultados numéri-cos, la interpretación para las dos líneas y las dos matrices se realiza de igual manera que para la anterior prueba. El valor de *Adjusted Squared Multiple R*, indica que aproximadamente el 99% de la varianza fue explicado por las variables dependientes (HORLUZ Y NITRO). Con *Std Coef* se puede observar cual de las variables influye más, horas luz (0.797) influye más que nitrógeno (0.449).

Conclusiones: Las variables hora luz y nitrógeno explican con gran certeza el crecimiento de las plantas. La hipótesis nula que pertenece al *F-ratio* y el valor de *P*, muestra que los coeficientes de las variables independientes son muy bajos (se desecha la hipótesis nula, no hay relación entre las dos variables). Por lo tanto, las dos variables son importantes para la explicación del crecimiento de las plantas.

Ejecutar: Stats ÷ Regression ÷ NonLinear ÷ Model/loss



Resultados/gráficos

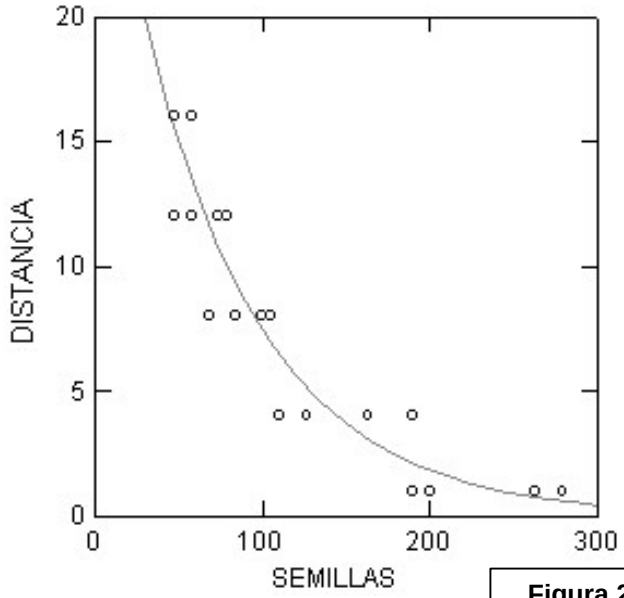


Figura 21

B1f. Regresión No - Lineal

Definición: Por medio de una ecuación, una variable independiente (eje X = SEMILLAS) se predice el comportamiento de una variable dependiente (eje X = DISTANCIA), ver Figura 21.

Objetivo: Determinar una ecuación para predecir la distribución de las semillas por medio de la distancia que existe entre el árbol padre y donde se encuentra la semilla.

Hipótesis: H_0 = No hay una relación entre el número de SEMILLAS y la DISTANCIA; H_a = Hay una relación significativa y marcada entre el número de SEMILLAS y la DISTAN-CIA.

Diseño: Se recolectaron semillas de alrededor de 20 árboles con trampas, ubicadas a 1, 4, 12 y 16 metros del árbol que produce semillas.

Interpretación: En el gráfico se muestra que la línea de la ecuación (Y=Exp(a+b*X)) des-cribe a los datos de la relación, porque pasa por la mitad de los extremos de los dados. De los resultados numéricos los más importantes son *Raw Rsquare* = 0.964 (lo que muestra

que hay una buena relación entre distancia y número de semillas), y los valores de A y B que se encuentran por debajo de *Estimate* (A=3.413 y B=-0.014), son importantes para reemplazar en la ecuación general. También es importante observar el error estándar de A y B que está por debajo de A.S.E. en este caso para A=0.051 y B=0.00074 (Figura 22).

Conclusiones: Las semillas se encuentran con mayor abundancia cerca a los productores de semillas. El modelo que mejor explica la distribución de datos entre el número de semillas y distancia, es Y=exp(A+B*X).

Resultados/numéricos

```
Data for the following results were selected according to:
      (ESPECIE$= "Caesalpinia")
 Iteration
 No.       Loss       A            B
   0 .176852D+04 .101000D-01-.102000D-01
   1 .158005D+04 .102010D+01-.130130D-01
   2 .618128D+03 .267135D+01-.146608D-01
   3 .540630D+03 .376523D+01-.131632D-01
   4 .806938D+02 .346661D+01-.137100D-01
   5 .700190D+02 .341403D+01-.139796D-01
   6 .700060D+02 .341273D+01-.139996D-01
   7 .700060D+02 .341272D+01-.139995D-01
   8 .700060D+02 .341272D+01-.139995D-01

Dependent variable is DISTANCIA
     Source   Sum-of-Squares    df  Mean-Square
 Regression        1853.994     2      926.997
   Residual          70.006    18        3.889

      Total        1924.000    20
Mean corrected      579.200    19

       Raw  R-square (1-Residual/Total)        =0.964
Mean corrected R-square (1-Residual/Corrected) =0.879
          R(observed vs predicted) square      =0.879
                                     Wald Confidence Interval
Parameter Estimate  A.S.E. Param/ASE Lower < 95%> Upper
 A         3.413    0.142  23.994    3.114        3.712
 B        -0.014    0.002  -6.811   -0.018       -0.010
```

Figura 22

Nota:

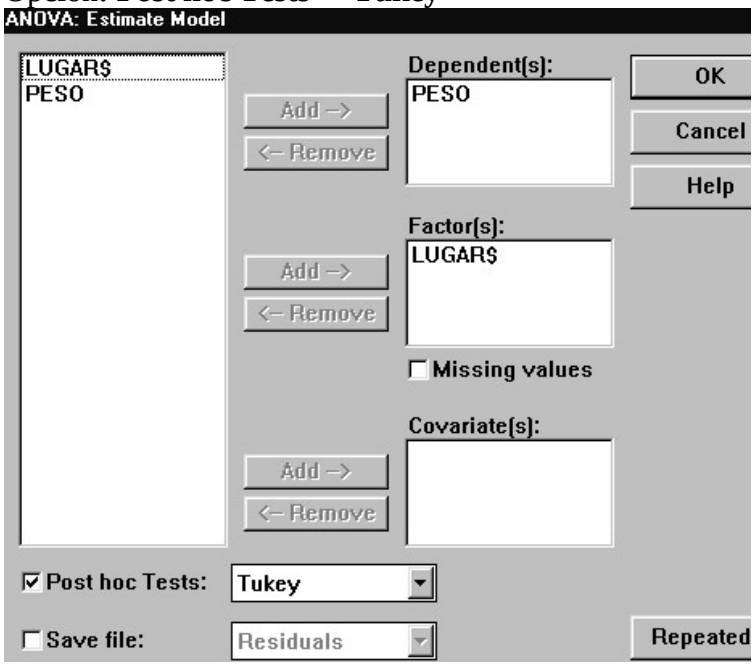
- No hay límite, en cuanto al número de tipos de curvas que puedan expresarse por ecuaciones matemáticas. Entre las curvas más conocidas están las Polinomiales, Exponenciales y Logarítmicas. Mientras que entre las ecuaciones están las lineales, cuadráticas y cúbicas (Cuadro IV-2).
- Algunas de las curvilíneas se pueden transformar por medio de logarítmicas y exponenciales; ejemplo: Una curva exponencial, dada por la fórmula $y=a*b^x$, se puede transformar en una ecuación logarítmica, dada por la fórmula log(y)=log(a)+x*log(b).

Cuadro IV-2: Algunas ecuaciones, de acuerdo a la curva distribuida por la variable independiente y dependiente.

Ecuación	Polinomial	Exponencial	Logarítmica
Lineal	Y=a+b*X	*e*^Y=a*X^b	Y=a+b*log(X)
Cuadrática	Y=a+b*X+c*X^2	Y=a*b^X	LogY=a+b*X
Cúbica	Y=a+b*X+c*X^2+d*X^3	Y=a*X^b	LogY=a+b*log(X)

- Ejemplo: Determinar el crecimiento del diámetro de un árbol a partir de un volumen de copa y los datos son: Volumen 22, 6, 93, 62, 84, 14, 52, 69, 99, 98, 41, 85 y 90; para crecimiento .36, .09, .67, .44, .72, .24, .33, .61, .64, .65, .47, .60 y 51. Los resultados indican por medio de R cuadrado, que hay una asociación entre Y(crecimiento) y X (copa), y su valor es 0.978. Esto quiere decir que el modelo lineal explica el 98% (redondeado). El modelo lineal es Crecimiento=a+b*volumen (igual a Y=a+b*X). Los valores de A = 0.162 y de B=0.005, y de estos el error estándar es A=0.052 y B=0.001.

Ejecutar:
Stats ÷ Analysis of Variance (ANOVA) ÷ Estimate Model
Opción**: Post hoc Tests ÷ Tukey**



Resultados/gráficos

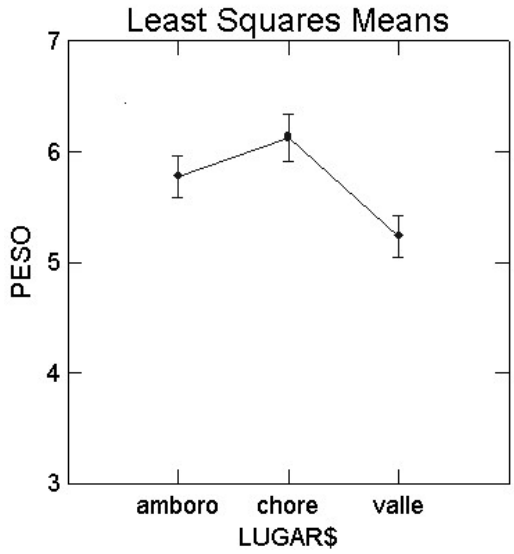


Figura 24

B1g. ANOVA de una Vía

Definición: Se aplica a más de dos grupos para comparar las medias y variaciones entre éllas. **Objetivo:** Determinar si los pesos de los armadillos (*Dasypus* sp.) son iguales o diferentes en distintos sitios (Figura 24).
Hipótesis: H_0 = las especies pesan igual en los tres bosques; H_1= los pesos son diferentes entre sitios. **Diseño:** Se pesaron 10 armadillos en el Chore, 12 en el Amboró y 12 en el Valle.
Interpretación: En los resultados gráficos, los pesos promedios de los armadillo de los tres sitios se muestran con su barra de error; cuando se observa horizontalmente las del Chore, se encuentran mas arriba y las del valle mas abajo; pero los errores estándar del Chore y de Amboró sus errores estándar se hallan solapados, y mientras, los errores estándar del valle no se solapan con lo de Chore y tampoco con el de Amboró (este es un indicador de las diferencias de promedios existentes). En los resultados numéricos *Multiple R* explica las correlaciones múltiples para la variable dependiente, mientras *Squared Multiple R* (raíz cuadrada del *Multiple R*) explica la variabilidad en porcentaje (24.6%) de la variable dependiente (Figura 25). En la matriz de análisis de varianzas el *F-Ratio & P* prueba la hipótesis alternativa de que hay diferencia de crecimiento de acuerdo a los lugares. En la segunda matriz se muestra las codificaciones de los lugares

Resultados/numéricos

```
>ANOVA>CATEGORY LUGAR$>COVAR>DEPEND PESO / TUKEY>ESTIMATE
Effects coding used for categorical variables in model.
Categorical values encountered during processing are:
LUGAR$ (3 levels)     amboro, chore, valle
DepVar: PESO N:34  Multiple R:0.496 Squared multiple R:0.246
                     Analysis of Variance
Source  Sum-of-Squares df  Mean-Square  F-ratio    P
LUGAR$     4.525        2      2.263     5.050     0.013
Error     13.890       31      0.448
**WARNING** Case5 is an outlier (Studentized Residual = 3.675
Durbin-Watson D Statistic    2.318
First Order Autocorrelation -0.248
COL/
ROW        LUGAR$
  1        amboro
  2        chore
  3        valle
Using least squares means.
Post Hoc test of PESO
Using model MSE of 0.448 with 31 df.
Matrix of pairwise mean differences:
                    1        2       3
              1    0.0
              2    0.355    0.0
              3   -0.542   -0.897   0.0
Tukey HSD Multiple Comparisons.
Matrix of pairwise comparison probabilities:
                    1        2       3
              1    1.000
              2    0.440    1.000
              3    0.134    0.010   1.000
```

Figura 25

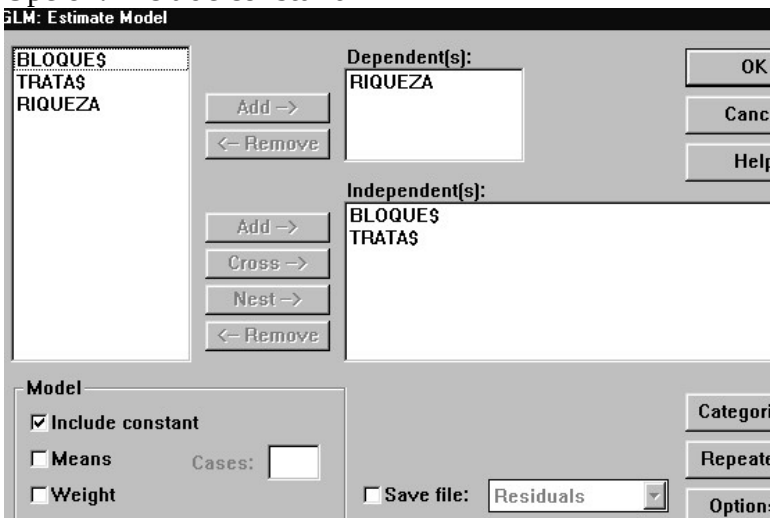
(1=amboró, 2=chore y 3=-valle). En la tercera matriz (*Matrix of Pairwise Mean Differences*) se muestra las diferencias de las medias. En la cuarta matriz (*Matrix of Pairwise Comparison Probabilities*) se muestran las probabilidades (de la diferencia de medias) por pares o entre dos sitios.

Conclusiones: La probabilidad en la matriz de *Analysis of Variance* indica que hay una diferencia significativa entre las poblaciones inferidas (P = 0.013), pero no dice entre cuáles; para ello se recurre a la matriz de probabilidades, la que muestra que entre 1 (amboró) y 2 (chore) no hay diferencia (P=0.44), entre 1 y 3 (valle) no hay diferencia (P=0.13) y entre 2 y 3 existe una diferencia significativa (P=0.01). Entonces se desecha la hipótesis nula y acepta la alternativa (Figura 25).

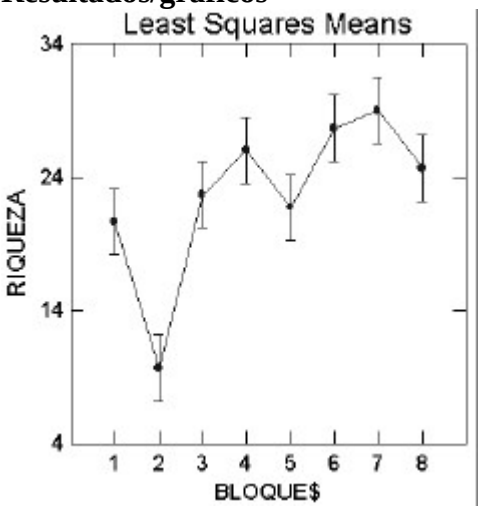
Nota:

- En las comparaciones las varianzas deben ser similares.
- Las mediciones deben tener una distribución normal o aproximada.

Ejecutar:**Stats ÷ General Linear Model (GLM) ÷ Estimate Model**
Opción**: Include constant**



Resultados/gráficos



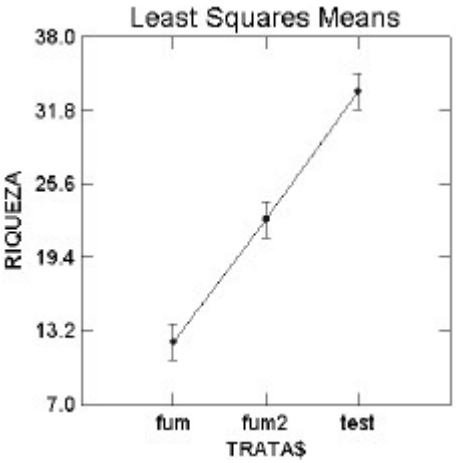


Figura 26

B1h. ANOVA de Dos Vías

Definición: compara las medias y las varianzas entre mas de dos factores (o variable independientes) contra mas de una variable dependiente.

Objetivo: Se quiere matar a los insectos barrenadores que afectan a las plántulas de mara con insecticida sin afectar mucho a los demás insectos.

Hipótesis: H_0**:** no hay diferencias entre los tres tratamientos, H_1**:** hay menos riqueza en los tratamientos fumigados, H_2**:** hay menos riqueza en las parcelas fumigadas con insecticida viejo.

Diseño: Probar si un insecticida viejo o nuevo mata más riqueza de insectos. Para esto se eligieron 8 plantaciones de mara y en cada una se localizaron tres sitos y a éstos se asignaron los tratamientos al azar: testigo (test), fumigado con insecticida viejo (fum) y fumigado con insecticida muevo (fum2). Después de dos días de la fumigación, en los tratamientos, se capturaron a los insectos sobrevivientes para hacer conteos de la riqueza de especies (Figura 26).

Interpretación: En los resultados gráficos se muestran: primer gráfico, que en los 8 bloques hay diferencia de medias con sus respectivos errores, de estas el bloque dos tiene menor riqueza de insectos y mientras el bloque 7 tiene mayor riqueza de especies de insectos; en el segundo gráfico se muestra que en las parcelas tratadas con

Resultados/numéricos

```
>MGLH

>MODEL RIQUEZA = CONSTANT + BLOQUE$+TRATA$ >ESTIMATE

Effects coding used for categorical variables in model.

Categorical values encountered during processing are:
BLOQUE$ (8 levels): 1, 2, 3, 4, 5, 6, 7, 8
TRATA$ (3 levels):  fum, fum2, test

Dep Var: RIQUEZA   N: 24   Multiple R: 0.953   Squared multiple R: 0.908

                              Analysis of Variance
Source              Sum-of-Squares   df  Mean-Square     F-ratio       P
BLOQUE$                    762.500     7      108.929       5.940       0.002
TRATA$                    1785.250     2      892.625      48.673       0.000
Error                      256.750    14       18.339
Durbin-Watson D Statistic     1.478
First Order Autocorrelation   0.255
```

Figura 27

insecticida viejo (fum) hay menor riqueza de insectos que en aquellas parcelas fumigadas con insecticida nuevo (fum2). En los resultados numéricos se describe al igual que en ANOVA de dos vías. En estos resultados las probabilidades de los bloques y tratamientos tienen una P menor a 0.05.

Conclusiones: Los resultados muestran que el insecticida viejo mata más riqueza de insectos que el nuevo, por lo cual se decide utilizar este último (Figura 27).

Nota:

- Se usa bloques para controlar la varianza que existe entre las plantaciones (esta varianza no interesa por que está excluida del efecto importante, "tratamiento").
- GLM y ANOVA sirven para hacer pruebas de análisis de varianzas; pero, el primero se usa cuando se realiza una replicación similar en todos los tratamientos (test=n=10; fum=n=10; y fum2=n10)y mientras el segundo se usa para replicaciones diferentes entre tratamientos (test=n=10; fum=n=12; y fum2=n12).
- Diseño de bloques completamente al azar: cuando los efectos de los bloques son significativos, quiere decir que la precisión del experimento ha aumentado debido al uso del diseño con aleatoriedad; si los efectos de los bloques son pequeños y no son significativos, el experimento no tiene éxito en reducir la varianza de las unidades y esto significa que las unidades experimentales eran homogéneas. Por lo tanto, los bloques se utilizan cuando hay sospecha de que hay factores extraños (local, tiempo, etc.) que afecten a los tratamientos.

B2. Transformación de Datos

B2a. Definición: Se transforman o alteran los valores de tal manera que queden en el mismo orden, con el objetivo de normalizar y disminuir la varianza de una muestra.
Para analizar los datos de una muestra con pruebas paramétricas se deben cumplir ciertos requisitos (ver flujo diagrama, Figura 8) y cuando no cumplen se debe tratar de hacer cumplir transformando a los datos; porque estas pruebas son más robustas o consistentes (la inferencia de una muestra a una población es mas buena) que las no paramétricas. En caso de no cumplir con la normalidad y varianza homogénea entre los comparados, como último recurso se utiliza a aquellos que trabajan con frecuencias como Ji - cuadrado, tablas de contingencia, pruebas de G, etc.

Existen varias alternativas para transformar de acuerdo a su distribución y dispersión, como se muestra en el siguiente Cuadro IV-3.

Cuadro IV-3: diferentes transformaciones de acuerdo a la distribución

TRANSFORMACION	UTILIZACION	EJEMPLO
Raíz Cuadrada	Se usan cuando los números son enteros y pequeños, los datos tienen una distribución de Poisson y cuando los eventos ocurren aleatoriamente en tiempo y espacio.	Abundancia de *Aspidosperma rigidum* que se encuentra en una parcela.
Logarítmica	Se usan cuando las varianzas o desviaciones estándar son proporcionales a los cuadrados de las medias de los tratamientos (esta transformación equilibra u homogéneas las varianzas) y se usa cuando la distribución se comporta de manera log - normal (esto se observa en conteos)	Número de insectos que se encuentran en una hoja, en la corriente de un río, en los troncos, etc.
Arcoseno	Se usa cuando los datos tienen una distribución bimodal y deben estar expresado en fracciones decimales o porcentajes (los valores de los datos deben comprender entre 0 y 1 antes de transformarlos).	Número de semillas que germinarn en una muestra de suelo.

También existen varias fórmulas para transformar los datos de una muestra de acuerdo a la distribución de la muestra, en el siguiente cuadro se sugieren algunas de las más comunes:

Cuadro IV-4. Algunas fórmulas, para transformar datos, que son utilizadas de acuerdo a las distribuciones más comunes (X = el nombre de la variable que se quiere transformar).

Distribuciones comunes	Fórmulas para transformar diferentes distribuciones	
	Funciones en SYSTAT	Fórmulas
J invertida	1/ (X)	$1 / X$
Muy sesgada a la derecha	LOG (X)	$\text{Log}(X)$
Moderadamente sesgada a la derecha	SQR (X)	$\sqrt{X}$
Moderadamente sesgada a la izquierda	-1/SQR(X)	$-1 / \sqrt{X}$
Muy sesgada a la izquierda	-1/LOG(X)	$-1 / \log(X)$
En forma de J	-1/(X)	$-1 / X$

Nota: X = valores de las muestras, LOG = logaritmo, SQR = raíz cuadrada.
En este cuadro aparecen seis fórmulas pero en realidad sólo existen tres por que las tres últimas son las distribuciones negativas o invertidas de las tres primeras. En las transformaciones con raíz cuadrada y logaritmos no deben existir valores negativos, y tampoco ceros, para evitar esto se suma 1 a todos los valores.

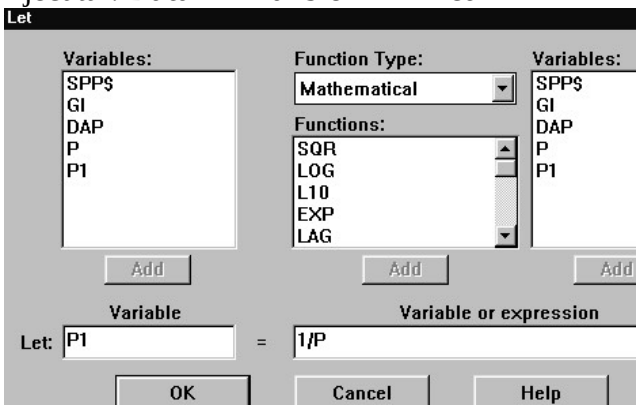
B2b. Transformación en SYSTAT

SYSTAT tiene diferentes fórmulas para transformar valores, lo que se puede encontrar en *Function Type* (como de matemáticas, estadística, etc.). Este programa, al transformar, reemplaza o crea una columna dependiendo de la ejecución: en *Let* se coloca un nuevo nombre de la columna (*Variable*), para este ejemplo sería P1, y este, es igual a una fórmula que transforma los datos de P1 (*Varaiable or expression*), ver Figura 28

Nota:

- Los valores transformados están ordenados por rangos; entonces, si un valor de una muestra A es mayor que la B, se mantendrá este orden después de la transformación. La transformación disminuye la influencia de los periféricos.
- Cuando se interpreta los datos, en el texto, se colocan los promedios provenientes de los datos no transformados y los valores de probabilidad se utilizadas de los datos transformados.

B3. Ejecutar: **Data ÷ Transform ÷ Let**

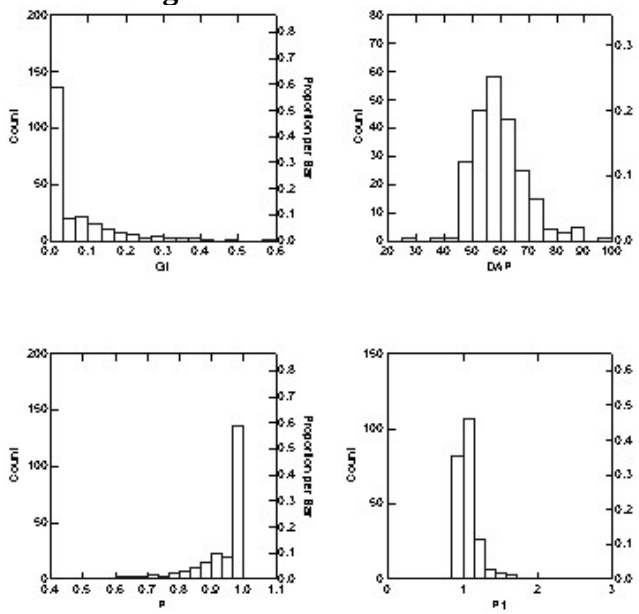


do

Ejecutar: **Graph ÷ Histogram (colocar en X-variable GI, DAD, P y P1). De estos salen los gráficos siguientes:**

Es

Resultados/gráficos



las

de

Figura 28

Pruebas No Paramétricas

Las pruebas no paramétrcias generalmente son homólogas de las pruebas paramétricas, por cual, cuanno se puede hacer análisis con las paramétricas (de acuerdo a los requisitos) se recurre a las no paramétricas (ver tabla). Estas pruebas trabajan con rangos, modas y medianas.

B3a. Mann-Whitney

una prueba homóloga de *t* para dos grupos. Aparece automática-mente cuando se ejecuta Kruskal - Wallis.

Definición: La prueba compara las medianas, modas y rangos de dos muestras que tengan igual o diferentes *n* (unidad de muestreo). Esta prueba se utiliza cuando las varianzas son heterogéneas y/o las variables son categóricas.

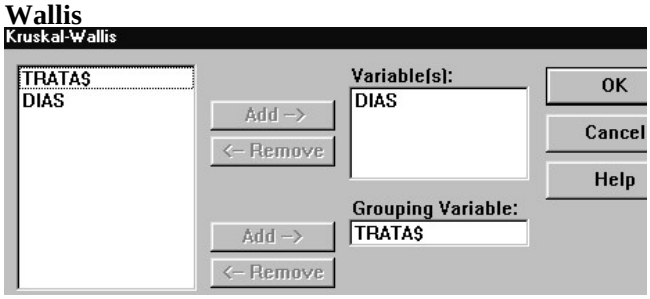
Objetivo: Las mariposas viven mayor tiempo alimentándose con azúcar o néctar.

Diseño: Se contó independientemente los días de sobrevivencia de mariposas alimentadas con solu-ción *azúcar* y con *néctar*.

Interpretación: En los resultados numéricos, en la matriz, se muestra a los grupos de azúcar y néctar (*group*) y sus valores de: número de muestras (*Count*) y la suma de rangos (*Rank Sum*); abajo de la matriz se muestra el valor de U = 0.0 y posteriormente la probabilidad que es 0.000, que es la misma que P<0.001, ver Figura 29.

Conclusión: De acuerdo a los rangos y la probabilidad, las dos poblaciones son diferentes. Dicho de otra manera, las mariposas alimentadas con néctar viven más días que aquellas alimentados con azúcar (P<0.0001).

Ejecutar: **Data ÷ No Parametric Tests ÷ Kruskal - Wallis**



Resultados/numéricos

```
>NPAR >KRUSKAL DIAS * TRATA$

Categorical values encountered during processing are:
TRATA$ (2 levels)
   Azucar, Nectar

Kruskal-Wallis One-Way Analysis of Variance for 34 cases
Dependent variable is DIAS
 Grouping variable is TRATA$

    Group        Count   Rank Sum
  Azucar           17     153.000
  Nectar           17     442.000
Mann-Whitney U test statistic = 0.0
Probability is        0.000
Chi-square approximation = 25.160   with 1 df
```

Figura 29

Ejecutar: **Data ÷ No Parametric Tests ÷ Wilcoxon**

Resultados/numéricos

```
>WILCOXON ANTES DESPUES

Wilcoxon Signed Ranks Test Results

Counts of differences (row variable greater than column)
                      ANTES      DESPUES
 ANTES                    0           13
 DESPUES                  1            0

Z = (Sum of signed ranks)/square root(sum of squared ranks)
                      ANTES      DESPUES
 ANTES                0.0
 DESPUES             -2.733        0.0

Two-sided probabilities using normal approximation
                      ANTES      DESPUES
 ANTES                1.000
 DESPUES              0.006        1.000
```

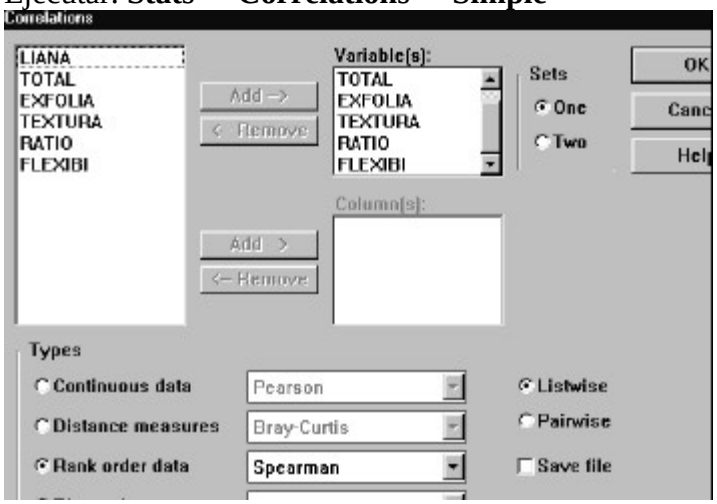
Figura 30

B3b. Prueba de Wilcoxon

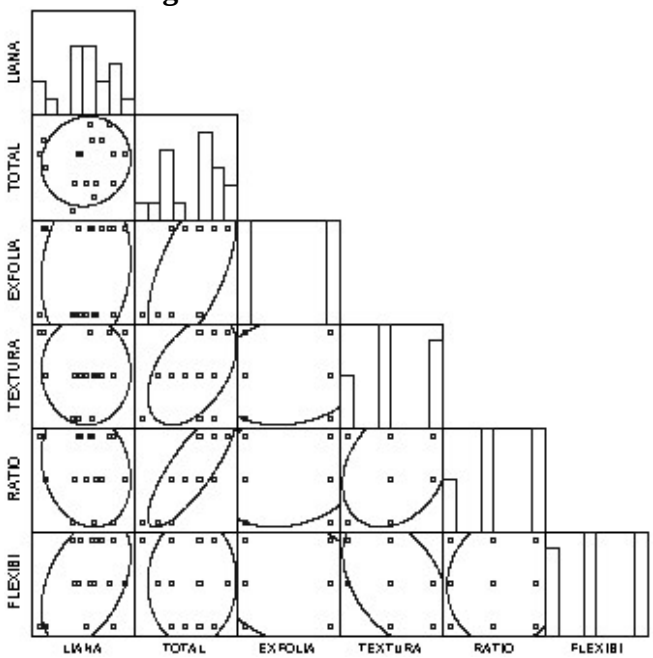
Esta prueba es homóloga de *t* pareada.

Definición: La prueba compara medianas, rangos y modas de dos muestras que son mediciones realizadas antes y después de un tratamiento; los requisitos son los mismos que para la anterior.
Objetivo: Determinar si la presión de agua que sube hacia la copa es influenciada por las lianas o no.
Diseño: El diseño es el mismo que para el ejemplo de *t* pareada.
Interpretación:
En la tercera matriz se muestra la probabilidad de la comparación de dos variables (el programa usa una aproximación de normal), donde el valor entre antes y después es 0.006 (Figura 30).
Conclusión: Existe una diferencia significativa de la presión de agua en los árboles antes y después de la corta de las lianas.

Ejecutar: **Stats ÷ Correlations ÷ Simple**



Resultados/gráficos



Resultados/numéricos

```
>CORR
>SPEARMAN LIANA TOTAL EXFOLIA TEXTURA RATIO FLEXIBI

Spearman correlation matrix
           LIANA  TOTAL  EXFOLIA TEXTURA RATIO FLEXIBI
 LIANA     1.000
 TOTAL     0.117  1.000
 EXFOLIA   0.205  0.627  1.000
 TEXTURA   0.076  0.458  0.186   1.000
 RATIO    -0.182  0.790  0.209   0.283   1.000
 FLEXIBI   0.339  0.065 -0.230  -0.451   0.072  1.000

Number of observations: 17
```

Figura 31

B3c. Correlación de Spearman

Es una prueba homóloga de la correlación de Pearson. Se utiliza cuando las muestras son pequeñas o cuando no tienen distribución normal (Cuadro IV-1).

Definición: Es una prueba que asocia a dos o más variables. Los datos originales deben estar en parejas, los valores absolutos son transformados en rangos y de éstos se calcula el coeficiente de correlación.

Objetivo: Determinar cuál de las características fisiológicas de las plantas, está asociada a la abundancia de las lianas.

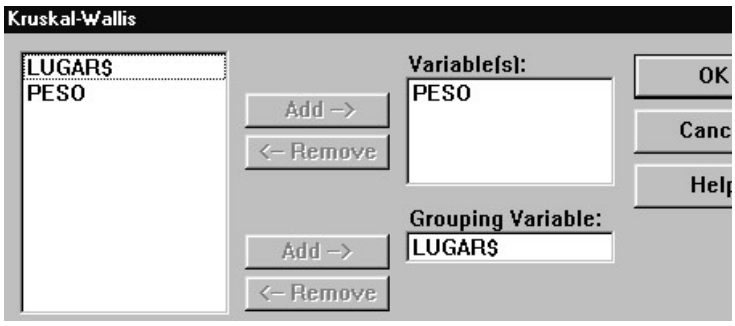
Diseño: En 17 árboles se estimó el grado de infestación con lianas (en porcentaje); posteriormente se registraron 4 características morfológicas y mecánicas (exfoliación, textura, radio y flexibilidad), y éstas se registraron en categorías de 1 a 4.

Interpretación: En los resultados gráficos se muestra las correlaciones entre todas las variables y cada relación se hallan visualizada dentro de un cuadrante y en éstas se hallan los puntos y un círculo o elipse, que muestran el grado de relación; ejemplo, entre TOTAL y RATIO hay una elipse inclinada hacia la derecha; esto indica que hay una relación fuerte y positiva (Figura 31). Los resultados numéricos muestran una matriz donde se encuentran los valores de *r* de Spearman de cada relación. Con el programa no se obtiene una matriz de probabilidades como en la correlación de Pearson; por el cual, los valores se buscan en una tabla de Pearson con grados de libertad (gl).

Conclusiones: Las conclusiones se realizan con los valores de *r* o se puede recurrir a tablas de Pearson que muestran las significancias con probabilidades. Existe una correlación fuerte entre Ratio y total y también entre exfoliación y total; y la correlación entre las demás variables (TEXTURA, LIANA y FLEXIBI) tienen valores de *r* por debajo de 0.60, por tanto existe una relación débil (Figura 31)

Nota: A diferencia de la correlación paramétrica, el coeficiente Spearman no tiene que tener una relación lineal, tampoco ninguna variable tiene que tener una distribución normal.

Ejecutar: **Data ÷ No Parametric Tests ÷ Kruskal - Wallis**



Resultados/numéricos

```
>NPAR >KRUSKAL PESO * LUGAR$

Categorical values encountered during processing are:
LUGAR$ (3 levels)  amboro, chore, valle

Kruskal-Wallis One-Way Analysis of Variance for 34 cases
Dependent variable is PESO
 Grouping variable is LUGAR$

    Group        Count   Rank Sum
  amboro           12     246.500
  chore            10     220.500
  valle            12     128.000

Kruskal-Wallis Test Statistic = 8.909
```

Figura 32

B3d. Kruskal-Wallis

Esta prueba es homóloga de la prueba de ANOVA de una vía.

Definición: Es una extensión de la prueba de Mann-Whitney. Se usa con más de dos grupos o tratamientos y trabaja con los rangos, medianas y modas.

Objetivo: Determinar si los pesos de los armadillos son diferentes entre tres sitios.

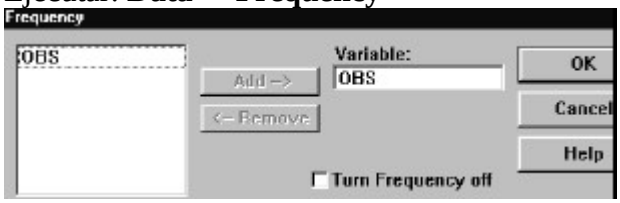
Hipótesis: H_0 = los pesos de los armadillos son similares en los tres lugares, Ha = los pesos entre sitios son diferentes.

Diseño: Se obtienen los pesos de armadillos aleatoriamente en los tres sitios.

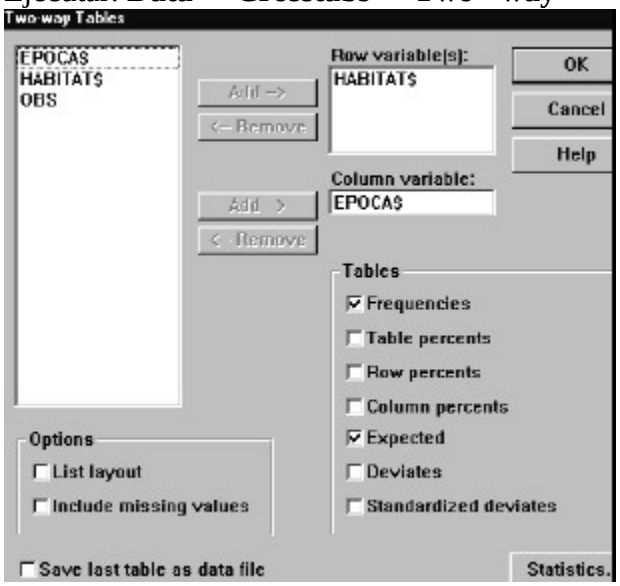
Interpretación: En los resultados numéricos se muestra una matriz donde se encuentran los datos estadísticos de los tres sitios, count y rank sum; por debajo se encuentran el estadístico de Kruskal Wallis y posteriormente la probabilidad, que es importante para interpretar las diferencias entre las muestras.

Conclusiones: Concluimos que los pesos promedios de los armadillos, comparados entre los tres lugares, hay diferencia significativa (*P* = 0.012), ver Figura 32.

Ejecutar: **Data ÷ Frequency**



Ejecutar: **Data ÷ Crosstabs ÷ Two - way**



Resultados

```
>FREQUENCY OBS >XTAB
>PRINT NONE/ FREQ EXPECT CHISQ
>TABULATE HABITAT$ * EPOCA$
   Case frequencies determined by value of variable OBS.
Frequencies
 HABITAT$ (rows) by EPOCA$ (columns)
         humeda   seca     Total
        +---------------+
     ba |    20     18  |     38
     bb |     8     15  |     23
     bs |    18     30  |     48
        +---------------+
 Total       46     63       109
Expected values
 HABITAT$ (rows) by EPOCA$ (columns)
           humeda    seca
         +-----------------+
      ba | 16.037  21.963  |
      bb |  9.706  13.294  |
      bs | 20.257  27.743  |
         +-----------------+
Test statistic              Value     df     Prob
  Pearson Chi-square        2.649   2.000   0.266
```

Figura 33

B3e. Ji - Cuadrado (X^2)

Definición: Analiza datos que son conteos del número de sujetos en una muestra que caen en categorías diferentes. Explica si existe una asociación entre dos factores en una población.

Objetivo: Determinar si los ciervos prefieren ciertos hábitats durante las épocas del año.

Diseño: Se recorrieron sendas registrando ciervos en los tres hábitats durante las dos épocas (con igual esfuerzo).

Hipótesis: H_0 = Los ciervos prefieren todos los hábitats en las dos épocas. Ha = Los ciervos prefieren algunas de los hábitats durante las épocas del año.

Interpretación: En la primera matriz se muestran las frecuencias tabuladas en los tres hábitats durante las dos épocas y las sumas totales. En la siguiente matriz se muestran los valores esperados de los tres hábitats durante las dos épocas. Finalmente se muestran los valores estadísticos calculados, el valor de X^2, grados de libertad y la probabilidad (0.266).

Conclusiones: Los ciervos prefieren de igual manera los hábitats en las diferentes épocas (X^2 = 2.649, gl = 2, P = 0.266), Figura 33.

Nota:

- Cuando hay más de un valor esperado menor a 5, las resultados no son adecuados. La prueba no asume ninguna distribución, por lo cual se pueden usar en diferentes situaciones. Además, recomendamos que se debe utilizar como un último recurso.
- X^2 se usa para la asociación entre dos factores y ellos son: Bondad de ajuste; Pruebas para asociación (a, No hay una relación entre el color de ojos y la dominancia de los individuos en grupo social; b, No hay una relación entre el color de las flores de *Tabebuia impetiginosa* y la humedad del suelo) y Pruebas de medias (Para frecuencias e intervalos de confianza, para una sola variable, se usa **Data ÷ Crosstabs ÷ One - way** y para frecuencias, porcentajes y medidas de asociación, de dos factores, se usa **Data ÷ Crosstabs ÷ Two – way).**

ANEXO 1

FORMA DE COLOCAR DATOS EN PLANILLAS ELECTRONICAS PARA ANALIZAR CON DIFERENTES PRUEBAS UTILIZANDO LOS PROGRAMAS SYSTAT Y JMP

PRUEBAS PARAMETRICAS

t de grupos

Tratamientos

GRUPO$	Noind
F	7
F	6
F	9
F	6
F	6
F	12
F	12
F	15
O	8
O	5
O	3
O	5
O	17
O	4
O	4

ANOVA DE 1 VIA

Un factor

GRUPO$	Noind
F	9
F	9
F	17
F	15
F	12
F	3
F	20
O	15
O	3
O	20
O	18
O	19
O	7
T	4
T	16
T	8
T	8
T	17
T	2
T	19

ANOVA DE 2 VIAS

Dos factores

BLOQUE$	TRATA$	Noind
I	I	82
I	II	67
I	III	65
2	I	77
2	II	68
2	III	72
3	I	62
3	II	66
3	III	90
4	I	87
4	II	72
4	III	73
5	I	82
5	II	84
5	III	70

t Pareada

2 muestras pareadas

ANTES	DESPUES
42	43
46	77
44	85
70	73
51	40
47	74
52	77
51	50
63	70
62	73
50	65
53	49
45	73
69	50
49	61

CORRELACION

asociación

LARGO	ANCHO
86	21
89	20
76	28
81	27
65	28
72	26
80	49
84	49
68	48
66	55
80	34
84	25
63	21
78	21
73	54

REGRESION

Predicción

DAP	DOSEL
81	82
76	98
48	70
81	98
60	87
87	76
97	60
91	80
45	58
85	75
58	83
88	68
51	52
60	58
93	78

Ejecución en SY	Stats, t-test, two group		Stats, Analy...(ANOVA), Estimate ...		Stats, Gener...(GLM), Estimate ...			Stats, t-test, paried		Stats, correlations, simple		Stats, regression, linear	
Muestras	*indep*		*indep*		*indep*	*indep*		*depen*	*depen*	*depen*	*depen*	*indep*	*depen*
Variables=	*categ*	*interva*	*categ*	*interva*	*categ*	*categ*	*interva*	*interva*	*interva*	*interva*	*interva*	*interv*	*interv*

Cuando los datos no tienen una distribución normal, sus varianzas son heterogeneas, ...etc. se recurre a las pruebas no paramétricas que son análogas a las pruebas paramétricas, por debajo de cada prueba se citan:

PRUEBAS NO PARAMETRICAS

Mann Whitney | **Kruskal Wallis** | **Friedman?** | **Wilcoxon** | **Spearman**

ANEXO 2
INTERPRETACION DEL VALOR DE PROBABILIDAD

Para interpretar los valores de las probabilidades no existen definiciones estándares, sin embargo sugerimos algunos:

Valor de P (probabilidad)	Interpretación
$P < 0.01$	Muy marcada en contra de la Ho
$0.01 \leq P < 0.05$	Evidencia moderada en contra de la Ho
$0.05 \leq P < 0.10$	Evidencia indicativa en contra de la Ho
$0.10 \geq P$	Poca o ninguna evidencia en contra de la Ho

Ejemplos de interpretación

Prueba de *t*

Existía evidencia de que las raíces de las plantas fertilizadas fueron más largas que las testigos ($P < 0.05$)

Se observó marcada evidencia que al añadir abono, el crecimiento de las raíces aumentó ($P < 0.001$)

ANOVA
Existía evidencia de un efecto de luz y la tasa de transpiración ($P < 0.05$)

Correlación

Existía una marcada correlación entre diámetro y altura ($r = 0.91$) que fue significativa a $P < 0.05$.

La correlación entre peso y altura no fue significativa ($P > 0.05$).

Regresión

Existía marcada evidencia ($P < 0.001$) de una débil asociación lineal entre contenido de agua y contenido de arena ($r^2 = 0.38$).

Existía evidencia ($P < 0.05$) de una fuerte asociación lineal entre contenido de carbono y altitud ($r^2 = 0.88$).

Comparaciones *A Posteriori*

Esta parte trata de los procedimientos de comparaciones múltiples. Se usan para enterarse las probabilidades entre tratamientos múltiples con el fin de determinar con exactitud cuáles son los niveles específicos del factor que producen diferencias significativas. Hay un montón de pruebas

de probabilidades que pertenecen a SYSTAT, aquí se detalla brevemente algunos de los más útiles o comunes.

T-Test

Las dos pruebas proveen protección para pruebas múltiples. La única diferencia es la manera de calcular la probabilidad.

Bonferroni: multiplica la probabilidad por el número de pruebas.

Dunn-Sidak: la probabilidad de n pruebas independientes se calcula como $1-(1-p)n$.

ANOVA y GLM

Scheffé: muy conservador

Bonferroni: cuando el número de comparaciones es pequeño, quizás sea, esta prueba más sensible.

Tukey: Cuando el número de comparaciones es muy grande, es demasiado fácil encontrar una diferencia, entre los comparados.

Fishers LSD: no ofrece nada de protección

Correlación

Bonferroni: para pruebas múltiples

Uncorrected: se asocia con una coeficiente de correlación singular. Se usa cuando se selecciona una correlación específica. Se pueden obtener las probabilidades solamente con la prueba de Pearson. Si se compara los resultados de ambos, se ve que hay diferencias entre las probabilidades.