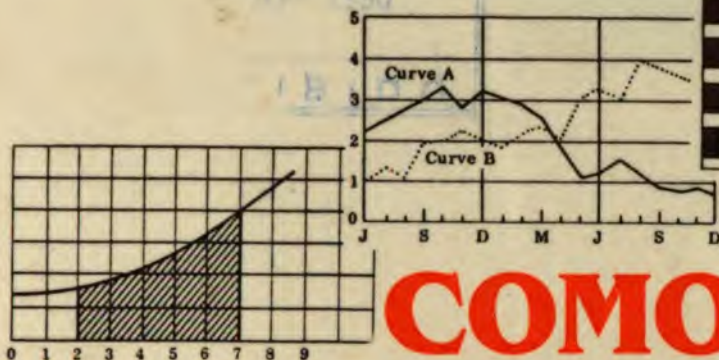
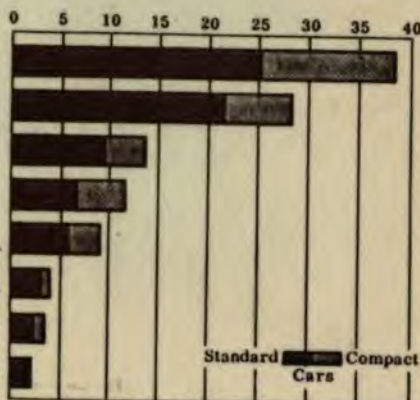
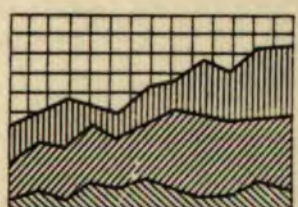


V
D

IICA



COMO USAR EL PAQUETE ESTADISTICO STATPAK



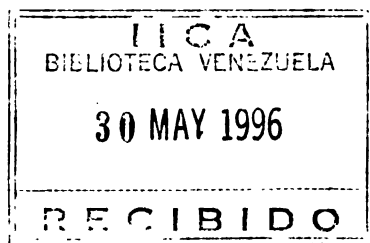
Y

Serie: Publicaciones Misceláneas
N° A4/UY-86-004
ISSN 0534-5391

MONTEVIDEO, URUGUAY

IICA





IICA-CIDIA

Serie Publicaciones Misceláneas
No. A4/UY-86-004
ISSN 0534-5391

COMO USAR EL PAQUETE ESTADISTICO STATPAK*

Jorge FRANCO DURAN **

Montevideo, Uruguay
Noviembre 1986

* STATPAK es Marca Registrada de Northwest Analytical Inc., de Portland, Oregon, USA.

** Ingeniero Agrónomo, MSc., Estadístico y Biometrista.

Esta publicación tiene la autorización correspondiente para mencionar el nombre de la firma Northwest Analytical Inc. y del producto STATPAK, conforme nota del Sr. Louis Halvorsen de fecha 2 de diciembre de 1986.

00001063

Franco Duran, Jorge

Como usar el paquete estadístico
STATPAK/Montevideo, Uruguay, IICA, no-
viembre de 1986.

202 p. Serie Publicaciones Misceláneas
N°A4/UY-86-004.

1. ESTADISTICA 2. COMPUTACION
 3. MENTALERIA

ISSN-0534-5391

CDD 519

PRESENTACION

La publicación de la guía de apoyo "COMO USAR EL PAQUETE ESTADISTICO STATPAK", elaborada por Jorge Franco Durán, tiene el propósito de servir de material de referencia para los investigadores agropecuarios, usuarios de esta mentalería de aplicación frecuente a los problemas agropecuarios.

El Proyecto de la Oficina en Uruguay del Programa II "Generación y Transferencia de Tecnología" pone a disposición de los investigadores y de otros usuarios el presente material, en la convicción que será de utilidad para su trabajo teniendo en cuenta la habilidad didáctica con que fuera escrito.

Horacio H. Stagno
Especialista en Programación de
la Investigación y Transferencia de
Tecnología Agropecuaria

B1-009401

11CA
PM 7:47.
86-00-1

INDICE

	<u>Pág.</u>
I. INTRODUCCION	1
II. CARACTERISTICAS GENERALES DEL STATPAK	3
2.1 Qué es el STATPAK	3
2.2 Algunas definiciones	4
2.3 Nombres	5
2.4 Programas disponibles en STATPAK	9
2.5 Instrucciones comunes a todos los programas	15
III. ORGANIZACION DEL TRABAJO CON STATPAK	25
3.1 Fase I. Entrada al Sistema	26
3.2 Fase II. Archivos de Datos	26
3.3 Fase III. Análisis Estadístico	27
3.4 Fase IV. Grabación/Eliminación de archivos y/o salida del Sistema	28
IV. DESCRIPCION DE PROGRAMAS	32
4.1 Programas para el manejo de archivos	33
4.2 Programas para generación de número aleatorios y manejo de distribuciones probabilísticas	77
4.3 Programas para la descripción estadística de muestras y poblaciones	97
4.4 Pruebas de hipótesis que utilizan la distribución "T de Student"	116
4.5 Programas para análisis de Regresión y Correlación	131
4.6 Programas para Graficación	173
4.7 Programas para el análisis de diseños experimentales	181

I. INTRODUCCION

La división del trabajo intelectual y la especialización actual de la ciencia hace cada vez más difícil encontrar un profesional que maneje múltiples disciplinas, se crea así la necesidad del trabajo interdisciplinario para la solución de los problemas científicos y se genera la posibilidad de que grupos especializados de profesionales produzcan medios que faciliten el trabajo a sus pares de otras áreas.

En la opinión de quien escribe este es el caso de los sistemas o paquetes de programas de computación que, elaborados por especialistas en programación y análisis de sistemas de cómputo, pueden ser utilizados por investigadores de otras áreas sin requerir ningún tipo de profundización en los conceptos teórico - prácticos de la programación de computadoras.

Así, independientemente de que la posibilidad de crear algo mejor, de más utilidad o simplemente diferente sigue abierta para quien enfrente el problema, este escrito pretende mostrar a investigadores del área agropecuaria algunos caminos para el procesamiento de sus datos experimentales mediante la utilización del paquete de programas Statpak, producido por Northwest Analytical, Inc. de Portland, Oregon, U. S. A.

La utilización de las computadoras personales como herramientas para el procesamiento de datos experimentales se puede dividir en dos etapas:

- (1) El manejo de archivos de datos que contengan, ordenada y analizable, la información recolectada por los investigadores o personal auxiliar en sus planillas y libros de campo, y
- (2) El análisis propiamente dicho, el cual usualmente contempla la aplicación de alguna técnica de la estadística descriptiva y/o inferencial.

En la primera etapa el Statpak provee facilidades para creación (directa o de otros sistemas como bases de datos, Lotus, etc.), modificación, transformación y en general, preparación de conjuntos de datos para su análisis. Respecto a la segunda etapa, posee programas que procesan estadísticamente los datos según las instrucciones dadas por el usuario.

Lo nuevo de este escrito es que trata de mezclar los aspectos mecánicos con la solución de ejemplos de aplicación y la discusión de algunas metodologías estadísticas, procurando ser una guía paso a paso y sin pretender reemplazar ni al manual de usuario de Statpak ni a un buen libro de métodos estadísticos ni, desde luego, el talento, la imaginación y la constancia de un buen investigador.

Se presentan tres capítulos, además de esta breve introducción:

- II. Definiciones generales y conceptualización.
- III. El método general de trabajo con Statpak.
- IV. Descripción de programas, presentación, solución y discusión de ejemplos.

Si logramos que el uso de esta herramienta permita a los investigadores tener un poco más de confianza en los métodos estadísticos, de precisión en sus conclusiones y de tiempo para pensar en las soluciones a los problemas de sus áreas específicas de trabajo, el tiempo y el esfuerzo gastados en producir el manual, estarán justificados.

II. CARACTERISTICAS GENERALES DEL STATPAK

Objetivos del Capítulo:

- Introducir al contenido general del STATPAK: qué es y para qué sirve?
- Familiarizar al lector con definiciones importantes.
- Enseñar a identificar archivos.
- Introducir a los programas disponibles en el STATPAK y a las áreas de análisis que contiene.

2.1 Qué es el STATPAK

El Statpak es un paquete de programas para el análisis estadístico de datos en microcomputadoras, elaborado en lenguaje de programación BASIC, que puede ser utilizado por personas que no tienen conocimientos de computación.

La idea fundamental es que un usuario del análisis estadístico pueda formar sus archivos de datos, corregirlos, transformarlos, unirlos, separarlos, aumentarlos, disminuirlos, etc. y posteriormente analizarlos estadísticamente y tener sus resultados en la pantalla para observación rápida, en papel impreso para un análisis más detallado, o guardados en un disco de trabajo para utilizarlos cuantas veces quiera.

Las instrucciones para utilizar el STATPAK son sencillas palabras en inglés, idioma de sus autores, que van siendo requeridas por el mismo paquete a medida que el trabajo se ejecuta y con las cuales el usuario se familiariza rápidamente.

En este manual abreviado se presentan las instrucciones requeridas para utilizar el paquete en los análisis más frecuentemente empleados por investigadores en el campo agropecuario. Su utilidad está directamente relacionada con el ejercicio que cada usuario tenga sobre la computadora misma, que es la mejor forma de adquirir habilidad en el manejo de cualquier herramienta. Se recomienda desarrollar muchos ejercicios y problemas de práctica. La posible diferencia de resultados entre la microcomputadora y el texto que tiene el ejemplo de práctica estimulará el análisis del uso y mostrará la propia potencia del STATPAK.

2.2 Algunas definiciones

(1) Archivo; Archivo de Datos (DATA FILE). Es el conjunto de datos que va a ser objeto de análisis. Es un arreglo matricial en hileras y columnas, de manera que cada dato está completamente identificado por su ubicación.

(2) Columnas (COLUMNS) e Hileras (ROWS). Puesto que se trata de análisis estadístico los archivos se organizan de tal forma que cada hilera es una observación (caso, o individuo) y cada columna es una variable (característica medida).

(3) Campos de Datos (DATA FIELDS). Son las variables (columnas) cuya definición o identificación es requerida por un programa, como mínimo, para poder trabajar. Una regresión, por ejemplo, precisa al menos dos campos de datos, uno para la variable X y otro para la variable Y.

La razón de que haya dos definiciones de las variables es que un archivo puede tener cualquier número de variables pero los análisis se hacen para cada una, o cada grupo, separadamente. El usuario indica qué variables entran al análisis en lo que se denomina una "asignación de campos".

(4) Procedimiento; Programa. Es cada uno de los programas que realizan las diversas funciones que posee el STATPAK (el programa SPLICE permite corregir archivos, el MLINREG hacer regresiones múltiples, etc.)

2.2.1 Ejemplo: En un experimento con caña de azúcar se midieron tres variables: porcentaje de azúcar, porcentaje de fibra y porcentaje de sacarosa. Se quiere estudiar la relación entre porcentaje de fibra y porcentaje de azúcar y sacarosa con el fin de predecir estas dos últimas a partir del porcentaje de fibra que es el más sencillo de medir. Los datos son:

Tabla Ejemplo 2.1. Caña de Azúcar: Porcentajes de Azúcar, Fibra y Sacarosa

Observaciones	V A R I A B L E S		
	(1) Porc. Azúcar	(2) Porc. Fibra	(3) Porc. Sacarosa
1	9.04	15.30	12.29
2	8.76	15.56	12.04
3	8.08	16.71	11.21
4	7.74	18.70	11.14
5	7.18	17.63	10.27
6	7.00	19.49	10.11
7	7.90	17.75	10.97
8	8.15	16.88	11.15
9	7.66	16.97	10.64
	(Y1)	(X)	(Y2)

Esta forma natural de presentar los datos es la forma en que trabaja STATPAK: un archivo de datos con 9 hileras (observaciones) y 3 columnas (variables).

Un programa sencillo que puede hacer el análisis de regresión lineal simple es el llamado "POLYREG". Al señalar a STATPAK que se quiere trabajar con ese programa en este archivo el SP indicará:

```
2 DATA FIELD (S)
3 FILE COLUMNS IN (nombre del archivo)
```

Esto es, que el programa solo precisa dos variables (DATA FIELDS) y en el archivo hay 3 (FILE COLUMNS). En seguida preguntará cuáles entran al análisis, correspondiendo la primera a X y la segunda a Y:

```
FIELDS ASSIGNMENT(S) ([CR] = 1:1): _
```

Si se le contesta [:2 1] entran al análisis como variable independiente (X) por ciento de fibra y como dependiente (Y) por ciento de azúcar (ver Cuadros 2.1 y 2.2)

Nota muy importante: Todos los procedimientos trabajan con esta organización de los archivos, excepto los de diseño experimental, que requieren formatos especiales y los cuales se tratarán más adelante.

2.3 Nombres

El usuario de STATPAK puede (y debe) poner nombres a sus archivos. Estos nombres constan de dos partes: el "nombre" propiamente dicho y la "extensión del nombre" separados por un punto

```
[nombre].[extensión]
ARROZ.DAT
```

El nombre deberá comenzar con una letra y tener un máximo de 8 caracteres (letras y/o números). Se recomienda que tenga algún sentido práctico para que facilite las cosas al usuario.

La extensión deberá tener solo tres letras y por comodidad se recomienda usar las siguientes extensiones al manejar los archivos:

```
[nombre].DAT: Archivo de DATOS
[nombre].PRM: Archivo de PARAMETROS
[nombre].FMT: Archivo de FORMATOS
[nombre].RSD: Archivo de RESIDUALES
[nombre].SRT: Archivo ORDENADO (con el
                procedimiento SORT)
```

2.3.1 Ejemplo: en el anterior ejemplo (ver Tabla Ejemplo 2.1) el archivo original se podría llamar AZUCAR.DAT. Si luego de la regresión se tiene un archivo con los errores residuales y valores estimados, este se podría llamar AZUCAR.RSD y si se ordenara en orden ascendente por la variable porcentaje de fibra, este archivo ordenado se llamaría AZUCAR.SRT.

Cuadro 2.1: Asignación de campos

---> POLYNOMIAL REGRESSION <---

Power of Polynomial (P)? 1

Residual File Name (<CR> to skip): AZUCAR.RSD

Calculate Durbin-Watson (Y/N)? N

Field 1 = Independent Variable (X)

Field 2 = Dependent Variable (Y)

Input File (Q=Quit, <CR>=WORK.DAT): AZUCAR.DAT

2 Data Field(s)

3 File Columns in AZUCAR.DAT

Field Assignment(s) (<CR>=1:1): 2 1



Cuadro 2.2 Análisis de regresión para porcentaje de azúcar (y)
 porcentaje de fibra (x)

--> PROG:POLYREG FILE:AZUCAR.DAT DATE:02-10-1986 ID:EJEMPLO DE LA ASIGNACION DE COLUMNAS
 PAGE 1

TERM	COEFFICIENT	STD. ERROR	T-STATISTIC	PART. CORR	CONTR. R-SQ
B 0	16.01868	1.406306	11.39061	---	---
B 1	-.4719897	8.201544E-02	-5.754888	0.8255	0.8255

	SUM SQ	DEG FR	MEAN SQ
DUE TO REGRESSION	2.91162	1	2.91162
ABOUT REGRESSION	.615402	7	8.791457E-02
TOTAL	3.527022	8	.4408777

R-SQUARED: .825518
 F-TEST: 33.11874
 CORRECTED R-SQUARED: .800592
 STD ERROR OF REG: .2965039


Cuadro 2.3 Archivos original, ordenado y de residuales para el ejemplo de CAÑA DE AZUCAR.

Cuadro 2.3.a

```

Rows to EOF:  9   Top Row Cols:  3   File: AZUCAR.DAT
 1:  9.04 15.30 12.29
 2:  8.76 15.56 12.04
 3:  8.08 16.71 11.21
 4:  7.74 18.20 11.14
 5:  7.18 17.63 10.27
 6:  7.00 19.49 10.11
 7:  7.90 17.25 10.97
 8:  8.15 16.88 11.15
 9:  7.66 16.92 10.64
-- END OF FILE --

```




Cuadro 2.3.b

```

Rows to EOF:  9   Top Row Cols: -3   File: AZUCAR.SRT
 1: 15.30  9.04 12.29
 2: 15.56  8.76 12.04
 3: 16.71  8.08 11.21
 4: 16.88  8.15 11.15
 5: 16.92  7.66 10.64
 6: 17.25  7.90 10.97
 7: 17.63  7.18 10.27
 8: 18.20  7.74 11.14
 9: 19.49  7.00 10.11
-- END OF FILE --

```




Cuadro 2.3.c

```

Rows to EOF:  9   Top Row Cols:  3   File: AZUCAR.RSD
 1:  9.04  8.797234  .2427654
 2:  8.76  8.674517  8.548355E-02
 3:  8.08  8.131730  -.0517292
 4:  7.74  7.428464  .3115358
 5:  7.18  7.697499  -.517499
 6:  7.00  6.819598  .1804023
 7:  7.90  7.876855  2.314568E-02
 8:  8.15  8.051491  9.850884E-02
 9:  7.66  8.032611  -.3726111
-- END OF FILE --

```



Las salidas obtenidas del STATPAK para este ejemplo se presentan en el cuadro 2.3

2.4 Programas disponibles en STATPAK

Aunque no se describen en este manual todos los programas que ofrece el STATPAK, se presenta una vista de aquellos de uso más frecuente con una breve descripción de cada uno. Para utilizar cualquiera de estos programas se recomienda consultar el manual original, que viene con los diskettes del programa o atreverse un poco y con mucho cuidado, llamando el programa de interés y dejándose guiar por las instrucciones que él mismo genera.

Los programas están agrupados por áreas de interés, de manera que haya una conexión lógica entre ellos. La lista de áreas y programas, con una breve referencia a su función, se da en las siguientes líneas:

2.4.1 Area A: MANEJO DE ARCHIVOS

- A: CONVERT: Permite el intercambio de archivos entre STATPAK y otros sistemas o programas.
- B: FORMOUT: Se usa en la escritura de archivos, por impresora, con un formato determinado (para hacer informes y/o reportes)
- C: FUNKEYS: Permite programar algunas teclas de la consola con instrucciones que se repiten muchas veces (facilita el manejo rápido del paquete)
- D: MISSDAT: Hace posible señalar con algún signo o número los datos faltantes en un archivo y/o reemplazar signos o números indicadores de datos faltantes (manejo de archivos con datos faltantes)
- E: ROTATE: Permite intercambiar hileras por columnas (trasponer la matriz) en el archivo de datos.
- F: SHUFFLE: Permite unir archivos horizontalmente por hilera, con base en los valores de una variable común a los dos archivos a unir
- * G: SORT: Hace posible reordenar las observaciones (hileras) de un archivo con base en los valores de una de sus variables
- * H: SPEED: Sirve para la creación, corrección, modificación, adición, unión, partición de archivos. Este programa es uno de los editores del STATPAK.

- * I: SPLICE: Es el otro editor del STATPAK; permite insertar, eliminar, extraer, reemplazar bloques y unir archivos en cualquier sentido.

2.4.2 Area B: FACILIDADES MATEMATICAS

- A: DESEASON: Permite la corrección de los datos de una serie de tiempo con un índice de estacionalidad (se usa en problemas de series de tiempo)
- * B: FUNGEN: Hace posible la construcción y evaluación de funciones matemáticas
- C: LAGDIF: Permite el retraso o adelanto de variables, reducción de todos los valores de una variable a una base común, expresando la reducción como simple diferencia o como porcentaje.
- * D: PLOT: Permite la graficación de conjuntos de datos o funciones matemáticas. Es uno de los programas de graficación del STATPAK, permite que el usuario haga sus propias especificaciones de forma.
- * E: QPLOT: Permite la graficación rápida de conjuntos de datos (Quick Plot Data). El programa por sí mismo hace las especificaciones de forma.
- F: RANKOUT: Genera rangos en orden ascendente o descendente para las variables indicadas por el usuario.
- * G: RANUM: Permite la generación de un conjunto de números aleatorios con distribución uniforme o normal.
- * H: SELECT: Permite dos tipos de acción: (1) selección de subconjuntos de datos a partir de un archivo, utilizando hasta diez criterios de selección; (2) transformación de datos, incluyendo operaciones que involucren diversas columnas del archivo.
- I: SMOOTH: Este programa permite "suavizar" curvas (curve smoothing) reduciendo el número de puntos, con técnicas como la de los promedios móviles, de bastante utilidad en problemas de series de tiempo.

2.4.3 Area C: PROBABILIDADES

- A: BAYES:** Este programa hace posible el uso de técnicas de bayes para el cálculo de probabilidades de eventos condicionados.
- B: COMBOS:** Permite el cálculo de combinaciones, permutaciones y factoriales.

2.4.4 Area D: CALCULO DE ESTADISTICAS PARA UNA VARIABLE

- * **A: DSTATS:** Este programa produce estadísticas descriptivas para una población o muestra.
- * **B: FREHIST:** Permite el análisis de frecuencias para variables numéricas (discretas o continuas) y hace el histograma respectivo.
- * **C: FREQUICK:** Este programa permite un análisis rápido de frecuencias para variables que toman valores numéricos discretos y/o alfabéticos o no numéricos como nombres, sexo (M o F), variedades (IRRI-12, Blue Bell).
- D: GENMEAN:** Este programa calcula las medias generalizadas de orden 1 a 10 para un conjunto de datos.
- E: DTSCORE:** Hace posible la estandarización de un conjunto de datos a una distribución normal [0.1] o (μ, σ^2) con μ y σ^2 definidos por el usuario.

2.4.5 Area E: FUNCIONES DE DISTRIBUCION DE PROBABILIDAD

- * **A: CONDIS 1:** Realiza cálculos en las siguientes funciones de distribución: normal bivariada, F de Snedecor, Gamma Incompleta, T de Student.
- * **B: CONDIS 2:** Realiza cálculos con las siguientes funciones de distribución: chi-cuadrado, exponencial, khagian-mazin, log-normal y normal.
- * **C: DISCDIS:** Hace cálculos con las siguientes funciones de distribución: binomial, hipergeométrica, binomial negativa, poisson.

2.4.6 Area F: REGRESION Y CORRELACION

- A: AUTOCO: Este programa permite el cálculo de los coeficientes de autorrelación y autocovarianza de cualquier orden.
- * B: CORREL: Hace el análisis de correlación lineal simple para parejas de variables.
- C: CROSSCO: Permite hacer análisis de autocorrelación y correlaciones cruzadas entre variables a diferentes grados de retraso (dependencias en series de tiempo)
- D: FOURIER: Este programa hace el análisis de fourier a una serie de tiempo. Calcula los coeficientes de fourier y el conjunto de datos transformados utilizando dichos coeficientes.
- * E: MLINREG: Este programa analiza modelos de regresión lineal múltiple. Permite la selección de variables mediante las opciones regular, forward - stepwise, backward stepwise o solución interactiva.
- * F: ONEVREG: Este programa efectúa el análisis de regresión para dos variables considerando cuatro modelos diferentes: lineal, exponencial, logarítmico y potencia.
- * G: POLYREG: Permite el análisis de regresión entre dos variables utilizando un modelo polinomial de cualquier orden.

2.4.7 Area G: PRUEBAS DE HIPOTESIS SOBRE PROMEDIOS Y ANALISIS DE CHI-CUADRADO

- A: CHISQ: Con este programa se obtiene el valor de chi-cuadrado calculado para prueba de Bartlett, diferencias entre proporciones y comparación entre frecuencias observadas y una misma o diferentes frecuencias esperadas.
- B: CONTAB: Este programa calcula el valor de chi-cuadrado para una tabla de contingencia de dos entradas, usado para probar la hipótesis de independencia de variables discretas.
- C: CROSSTAB: Este programa permite la elaboración de una tabla de clasificación de doble entrada.

- D: FISHER:** Permite hacer la prueba exacta de Fisher para tablas de contingencia 2 x 2.
- * E: MTESTS:** Con este programa se hacen pruebas de hipótesis que utilizan la distribución T de Student: correlación, comparación de dos poblaciones (pareadas o independientes) y correlación de Pearson para muestras pareadas.
- F: TABOUT:** Es un programa que complementa a CROSSTAB, formateando y enviando sus resultados a pantalla o impresión.

2.4.8 Area H: ESTADISTICA NO PARAMETRICA

- A: FRIEDMAN:** Este programa hace la prueba no paramétrica equivalente al análisis de varianza de 2 vías, para una variable medida en una escala al menos ordinal.
- B: KENDALL:** Con este programa se hacen pruebas de hipótesis sobre la independencia de dos variables con escala por lo menos ordinal.
- C: KOLSMIR:** Con este programa se hace la prueba no paramétrica para la bondad de ajuste de un conjunto de datos muestrales a una distribución teórica (Kolmogorov-Smirnov).
- D: KRUSWALL:** Este programa es equivalente al análisis de varianza completamente al azar tradicional, cuando las variables de interés están medidas con escalas ordinal o nominal (Kruskall-Wallis).
- E: MANWHIT:** Con este programa se puede hacer una prueba no paramétrica equivalente a la prueba de T para muestras independientes (Mann-Whitney).
- F: SPEARANK:** Permite calcular el coeficiente de correlación de Spearman entre dos conjuntos de rangos para probar hipótesis de independencia entre las variables "ranquedas" (ordenadas, según cierto criterio).
- G: WILCOSR:** Permite hacer una prueba no paramétrica equivalente a la prueba de T para muestras pareadas (propuesta por Wilcoxon).

2.4.9 Area I: DISEÑO EXPERIMENTAL

Con estos programas se puede calcular 1/:

- * A: ANOVA1: Diseño completamente al azar.
- * B: ANOVA1R: Diseño de bloques al azar.
- * C: ANOVA2: Factoriales A x B en diseño completamente al azar o bloques al azar con muestreo en las unidades experimentales.
- * D: ANOVA2R1: Parcelas divididas en diseño completamente al azar.
- * E: ANOVA2R2: Factorial A x B en diseño de bloques al azar o parcelas divididas en bloques al azar.
- * F: ANOVA3: Factorial A x B x C en diseño completamente al azar.
- * G: ANOVA3R1: Parcelas divididas en diseño completamente al azar en donde las parcelas principales están formadas por la combinación factorial de 2 factores.
- * H: ANOVA3R2: Análisis de varianza para un factorial A x B x C con medidas repetidas sobre dos tratamientos.

[1] Nota: Es muy importante notar que con una adecuada selección de sumas de cuadrados y grados de libertad, los programas anteriores sirven para analizar otro tipo de modelos experimentales.

2.5 Instrucciones comunes a todos los programas

Con el fin de desarrollar un sistema sencillo de interacción con el usuario, todos los programas de STATPAK tienen un mismo formato para preguntar:

- (1) El nombre del archivo que se debe analizar (INPUT FILE DECLARATION);
- (2) La asignación de campos del archivo de datos al programa (FIELD ASSIGNMENT);
- (3) A dónde deben enviarse los resultados (OUTPUT DIRECTION);
- y
- (4) Al salir del programa en cuestión qué hacer (PROGRAM EXIT).

A continuación se da una breve descripción de estas instrucciones, recordando que se requerirá usarlas permanentemente.

2.5.1 Input File Declaration

Forma: Input File (Q=Quit, CR = WORK.DAT [1]): .

Opciones	Acciones
a) Q <CR >	Se suspende la ejecución del programa.
b) <CR >	El programa asume que hay un archivo cuyo nombre es WORK.DAT en donde encontrará los datos que requiere.
c) [nombre de archivo] <CR >	El programa busca los datos en el archivo llamado [nombre de Archivo]. En caso de que el archivo definido no exista o no corresponda su forma con la requerida por el procedimiento aparecerá un mensaje de error y, de nuevo, la pregunta.

[1] Cuando ya se ha trabajado otro procedimiento STATPAK en el sitio de WORK.DAT aparece el nombre del último archivo utilizado. Este archivo es llamado el "archivo de entrada si no hay otra especificación" o "archivo de entrada por defecto" (Default Input File).

2.5.2 Field Assignment

Forma: [m] Data fields
 [n] File columns in [nombre de archivo]
 Field assignment(s) (CR = 1:1):_

Indica que el programa precisa m variables (columnas) y en el archivo hay n ($m \leq n$), pregunta luego cuáles columnas del archivo deben ser usadas por el programa.

Opciones	Acciones
a) <CR >	El Programa tomará las primeras m columnas del archivo, en orden, para hacer los análisis.
b) [Número de columna _número de col._... _número de col.]<CR >	Indica al programa los números de las columnas (variables) en el archivo, que deben ser usadas en el análisis, y el orden en que cada una entra. Muchos programas STATPAK requieren determinado orden en las variables para darles la función apropiada en el análisis, la asignación permite este tipo de definiciones. (ver el ejemplo de la Sección 2.2, cuadro 2.1)

2.5.3 Output Direction

Forma:

Output: S (screen), P (printer), F (file) or Q (uit)? _

Opciones [1]	Acciones
S < CR >	Los resultados del análisis aparecerán en la pantalla.
P < CR >	Imprime los resultados
F < CR >	Graba los resultados en un archivo, por cuyo nombre preguntará en el paso siguiente
Q < CR >	Termina, se dirige a preguntar por la salida del programa (Program Exit)

[1] Si la acción es S, P o F, la pregunta volverá a aparecer, hasta que la acción sea Q. El objeto de esto es permitir la revisión de resultados en la pantalla antes de hacer impresiones o grabaciones

2.5.4. Program Exit

Forma: Exit Command: CR = SP, R (erun), S (ystem),
or PROGRAM - .

Opciones	Acciones
CR	Sale del programa en cuestión y se dirige al programa SP apareciendo en pantalla la instrucción inicial (ver capítulo III, sección 3.1 del manual)
R CR	Vuelve al comienzo del programa en cuestión para iniciar otro análisis
S CR	Sale del STATPAK quedando en el sistema operativo de la computadora
[nombre de programa] [opción] CR	Sale del programa en cuestión y se dirige al procedimiento [nombre de programa] para hacer otro análisis. Si luego del [nombre de programa] y separada por un blanco se digita una O (opcional) el archivo de salida del programa en cuestión será el Input File del programa llamado.

Ejemplo: Sobre el conjunto de datos de caña de azúcar. Observe detenidamente y repita sobre la máquina las instrucciones señaladas con una flecha en las hojas denominadas "pantalla de instrucción" a medida que vayan apareciendo las preguntas respectivas. En las hojas denominadas "Resultado" observe el efecto de las instrucciones. Imagine, pruebe algunas otras instrucciones que se le ocurran, vea los resultados!!

Pantalla 2.1 : Instrucción

>>-----> S T A T P A K 3.1 <-----<<

Copyright (C) 1983 - NorthWest Analytical, Inc.
Portland, Oregon

COLUMN Mode active. Report Date: 02-13-1986

Enter: Program Name (or <CR> for File Aids Menu)
Program: POLYREG

Pantalla 2.2 : Instrucción


Power of Polynomial (P)? 1 

Residual File Name (<CR> to skip): AZUCAR.RSD 

Calculate Durbin-Watson (Y/N)? N 

Field 1 = Independent Variable (X)


Field 2 = Dependent Variable (Y)


Input File (Q=Quit, <CR>=WORK.DAT): AZUCAR.DAT 

2 Data Field(s)

3 File Columns in AZUCAR.DAT


Field Assignment(s) (<CR>=1:1): 2 1

Reading Data..... 

Output : S(creen), P(rinter), or Q(uit)? P 

--> MAKE PRINTER READY

Report ID: PRUEBA 

Output : S(creen), P(rinter), or Q(uit)?Q 

Exit Command: <CR>=SP, R(erun), S(ystem), or PROGRAM -- SORT 0 

Pantalla 2.3 : Resultado

--> PROG:POLYREG FILE:AZUCAR.DAT DATE:02-13-1986 ID:PRUEBA PAGE 1

TERM	COEFFICIENT	STD. ERROR	T-STATISTIC	PART. CORR	CONTR. R-SQ
B 0	16.01868	1.406306	11.39061	---	---
B 1	-.4719897	8.201544E-02	-5.754888	0.8255	0.8255


	SUM SQ	DEG FR	MEAN SQ
DUE TO REGRESSION	2.91162	1	2.91162
ABOUT REGRESSION	.615402	7	8.791457E-02
TOTAL	3.527022	8	.4408777

R-SQUARED: .825518 CORRECTED R-SQUARED: .800592
 F-TEST: 33.11874 STD ERROR OF REG: .2965039

Pantalla 2.4 : Instrucción


---> SORT UTILITY <---

of Columns in Output File? 3 

Data Column #1 is Sort Key 

Input File (Q=Quit, <CR>=AZUCAR.RSD):

of Columns in Output File? 3 

Data Column #1 is Sort Key 

Input File (Q=Quit, <CR>=AZUCAR.RSD):

3 Data Field(s) 

3 File Columns in AZUCAR.RSD 

Field Assignment (s) (<CR>=1:1): 2 1 3

Numeric or Alpha Sort Key (N,A)? N 

Ascending or Descending (A,D)? A 

Output File Name ('Q' to Quit): AZUCAR.SRT 

Determining File Size.....

9 records to sort.

Reading Data.....

Sorting.....


Writing.....

Done - 9 Records writ to Output.

Exit Command: <CR>=SP, R(erun), S(ystem), or PROGRAM -- QPLOT 0 

Pantalla 2.5 : Instrucción

---> QUICK PLOT <---



Maximum # of Y fields: 4 Data Field 1 = X-values
Data Fields 2 - 5 = Y-values 

How many Y fields to plot (<CR>=1,0=Quit)? 1

Input File (Q=Quit, <CR>=AZUCAR.SRT): 

3 File Columns in AZUCAR.SRT

(Enter 0 to Auto-Generate X)

File Field for X values (<CR>= 1): 1 File Field for Y 1 values (<CR>= 2): 3 Finding limits..... 

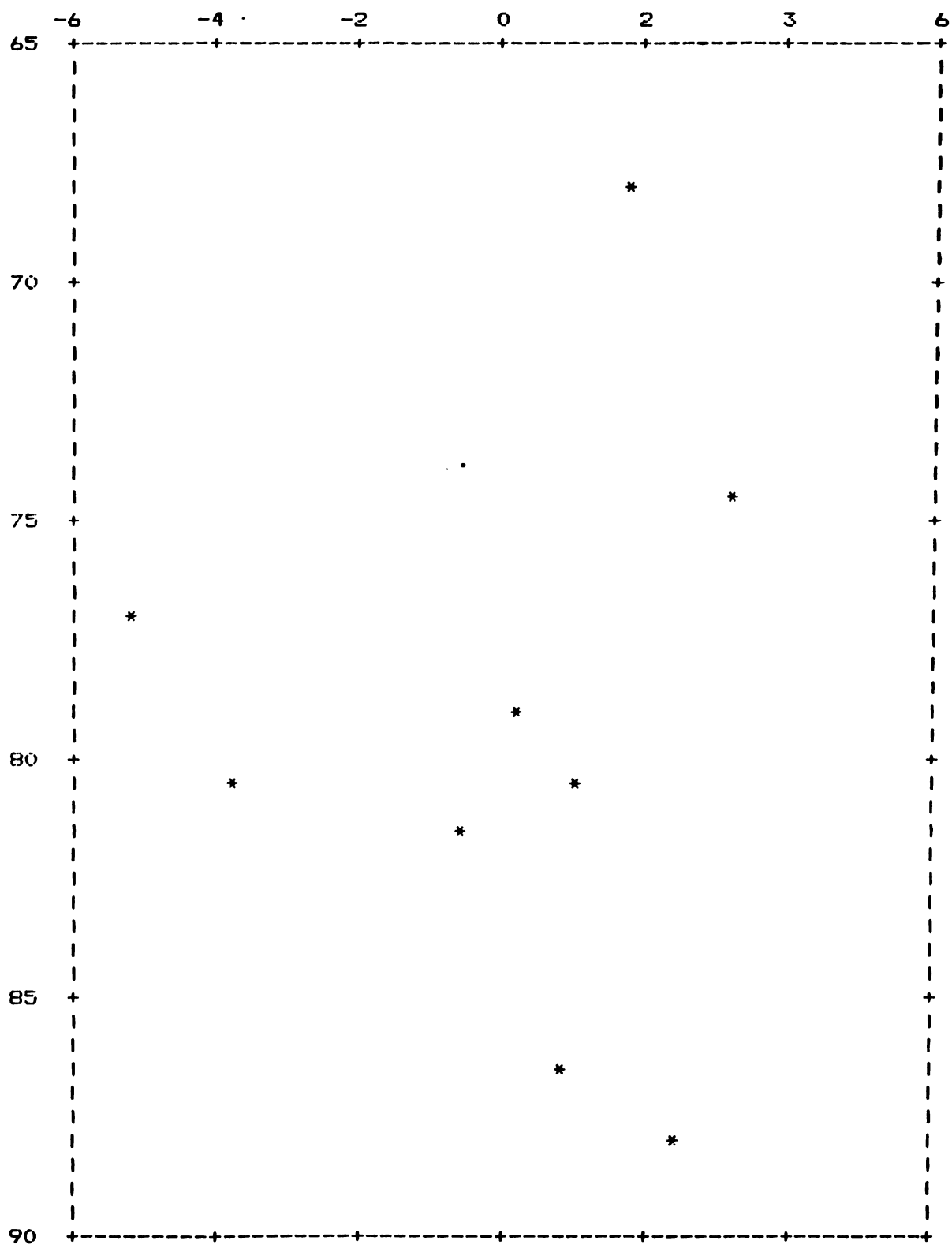
Output : S(screen), P(rinter), F(ile), or Q(uit)? P

---> MAKE PRINTER READY <---

Plot ID: PRUEBA Output : S(screen), P(rinter), F(ile), or Q(uit)? Q 

* .1 Y * .1 COLLISION MARK: ?

FILE: AZUCAR.SRT Columns: X = 1 (*) = 3



III. ORGANIZACION DEL TRABAJO CON STATPAK

Objetivos del capítulo:

Describir y ejemplificar el flujo de organización de una sesión de trabajo con STATPAK.

Se puede considera que una sesión de trabajo con STATPAK consta de cuatro fases:

- FASE I. Entrada al Sistema
- FASE II. Manejo de archivos de datos
- Fase III. Análisis estadístico
- Fase IV. Grabación y/o eliminación de archivos y salida del sistema

La Fase I requiere el conocimiento de algunas instrucciones para el manejo del sistema operativo de la computadora.

Las Fases II y III pueden interconectarse generando un "flujo de trabajo", en que un archivo creado por un procedimiento de análisis es utilizado por otro. Un ejemplo de este flujo se presentó en la sección 2.5, donde se hizo la gráfica de residuales para una regresión, a partir del archivo de valores esperados y residuales generados por el programa POLYREG.

La Fase IV se refiere a la biblioteca de datos que se sugiere mantener, utilizando discos blandos (diskettes) y a la eliminación, del disco duro, de archivos de trabajo que ocupan memoria innecesariamente y que obviamente estarán guardados en forma permanente en diskettes.

En lo que sigue se da una descripción de las diferentes fases, marcando las instrucciones requeridas para la correcta utilización del STATPAK.

3.1. FASE I. Entrada al Sistema

Luego de prender la máquina y dar las respuestas correctas a las preguntas de fecha, hora, etc. se estará dentro del sistema ope-

rativo de la computadora. Esta lo indica con la aparición en pantalla de la línea:

comp: C _

A la cual se debe contestar:

Usuario: C >CD\STATPAK

Comp.: C >

Usuario: C >BASICA SP

Comp.:

>>-----> S T A T P A K 3.1 <-----<<

Copyright (C) 1983 - NorthWest Analytical, Inc.
Portland, Oregon

COLUMN Mode active. Report Date: 02-12-1986

Enter: Program Name (or <CR> for File Aids Menu)
Program:

En este punto se está dentro del STATPAK y específicamente dentro de un programa llamado SP, que permite llamar cualquier otro y/o hacer algunas operaciones sobre archivos que ya existan y estén guardados en el disco duro y/o en diskettes.

3.2 FASE II. Archivos de Datos

Con esta fase comienza el trabajo STATPAK propiamente dicho. Es la etapa en que se preparan los archivos que se someterán a análisis. Las posibilidades que se dan en este punto son las siguientes:

- . Crear un archivo nuevo (entrando los datos por la consola)
- . Utilizar un archivo que ya está grabado en un diskette (drives A y/o B) o en el disco duro (drive C)
- . Utilizar partes de un archivo
- . Utilizar la unión de partes o de archivos completos
- . Transformar datos de un archivo

- . Hacer correcciones sobre un archivo
- . Ordenar las observaciones.
- . Seleccionar partes de archivos, etc..

Todas estas posibilidades y algunas otras, que no aparecen en la lista, pueden ser resueltas con los programas de STATPAK pertenecientes al área A: manejo de archivos y alguno del área B: facilidades matemáticas.

Es fundamental en esta etapa la noción de "Programa Editor". Un programa editor está construido para que el usuario pueda crear, corregir, transformar y hacer otro tipo de manipulaciones con archivos (de datos en el caso de interés o de cualquier tipo en el caso general). STATPAK posee dos programas editores: Speed (STATPAK ENHANCED EDITOR) y Splice (STATPAK LINE AND COLUMN EDITOR) diferenciados fundamentalmente en dos aspectos:

Speed (1) sirve para crear un archivo entrando los datos desde la consola y (2) ejecuta cualquier modificación directamente sobre el archivo que se está procesando.

Splice (1) trabaja sobre archivos ya grabados y (2) genera nuevos archivos con las modificaciones indicadas por el usuario dejando intacto el archivo que se está procesando.

Los programas que se describirán para esta fase II son los siguientes: Speed y Splice (Editores), Shuffle (Unión Horizontal de Archivos), Select (Transformación de Datos y/o Selección de Subconjuntos de Datos) y Sort (Ordenamiento Ascendente o Descendente).

3.3. FASE III. Análisis Estadístico

Una vez que se han preparado convenientemente los archivos de datos se procede a su análisis utilizando para ello los diferentes procedimientos. En la sección 2.4 se dio una breve descripción de los programas disponibles y se señalaron aquellos correspondientes a métodos estadísticos usuales en investigación agropecuaria. En el Capítulo IV se dará una descripción completa de cada uno de ellos.

Es importante que el usuario conozca la estructura que cada programa exige del archivo de datos para poder utilizarlo en esta fase de análisis, este conocimiento le dará las pautas para el manejo de archivos en la fase II. Esto es especialmente válido para los programas de análisis de diseños experimentales.

La etapa de análisis desde luego no se restringe al análisis de un archivo, con un procedimiento, sino que incluye cualquier posibilidad, como aplicar diferentes procedimientos al mismo o a diferentes archivos, como generar archivos con un procedimiento para analizarlos con otro, etc..

Al terminar con el último programa de análisis, el usuario tendrán en pantalla la instrucción de dirección de la salida de resultados:

OUTPUT: S(creen), P(rinter), F(ile), or Q(uit)?.

a la cual responderá Q <CR> para pasar a la instrucción de salida del programa:

Exit Command:<CR>= SP, R(erun), S(ystem), or Program.

con la cual se inicia la fase IV.

3.4 FASE IV. Grabación/Eliminación de archivos y/o Salida del Sistema

Terminado el análisis de los datos el usuario puede proceder en dos direcciones:

- (1) Eliminar del disco duro los archivos utilizados en la sesión y salir del sistema, en caso de que los archivos empleados ya estén grabados en diskettes,
- (2) Grabar en diskette los archivos creados y/o transformados en la sesión, eliminarlos del disco duro y salir del sistema. La razón para la eliminación de archivos del disco duro es mantener este disco tan limpio como sea posible para que otros usuarios tengan en sus sesiones de trabajo el máximo de capacidad disponible.

En cualquiera de los dos casos las instrucciones requeridas están en el programa "SP" al cual se accede con la opción <CR> del "Exit Command" (recuerde que la forma de esta instrucción es: Exit Command: <CR> = SP, R(erun), S(ystem), or Program .).

La computadora responderá a <CR> con el Menú de Facilidades para el manejo de archivos (FILE AIDS MENU) que se presenta en el cuadro

3.2. Este menú muestra las diferentes acciones que se pueden llevar a cabo, tres de las cuales se explican más adelante con pequeños ejemplos.

En el cuadro 3.3 se dan ejemplos sobre el uso de las instrucciones para Borrar (Delete), Renombrar (Rename) y Copiar (Copy) archivos. En el primer caso se borró el archivo AZUCAR.DAT que estaba grabado en el disco duro (Drive C, que es el que la máquina asume si no se le da ninguna especificación de Drive). En el segundo se renombró el archivo AZUCAR.DAT, que de ese momento en adelante pasa a llamarse CAÑA.DAT (estaba y queda en el disco duro). En el tercer caso se copió el archivo AZUCAR.DAT del Drive C a un diskette ubicado en el Drive A y se le dejó el mismo nombre (en ese momento AZUCAR.DAT está grabado en dos sitios, el disco duro y un diskette).

Se recomienda a los usuarios probar con estas y otras instrucciones usando archivos que no sean importantes, pero se aconseja tener

Cuadro 3.2 : Menú de ayuda para manejo de archivos

```
<-----<< FILE AIDS MENU >>----->          REPORT DATE: 02-17-1986

A lter Report Date
O utput to printer
S witch mode to ROW

C opy a file
R ename a file
D elete a file
L ist file names
E xchange Disk(s)

X ecute a STATPAK program

B ASIC command mode
Q uit (to system)

Select desired function (A,O,S,C,R,D,L,E,X,B,Q)?
```

Cuadro 3.3 : Instrucciones para: (a) borrar; (b) renombrar y (c) copiar archivos.

Cuadro 3.3.a

File to delete (<CR> to quit): AZUCAR.DAT
File AZUCAR.DAT deleted.

File to delete (<CR> to quit):

Cuadro 3.3.b

File to rename (<CR> to quit): AZUCAR.DAT
Rename it to (<CR> to quit): CAÑA.DAT
File AZUCAR.DAT renamed to CAÑA.DAT.

File to rename (<CR> to quit):

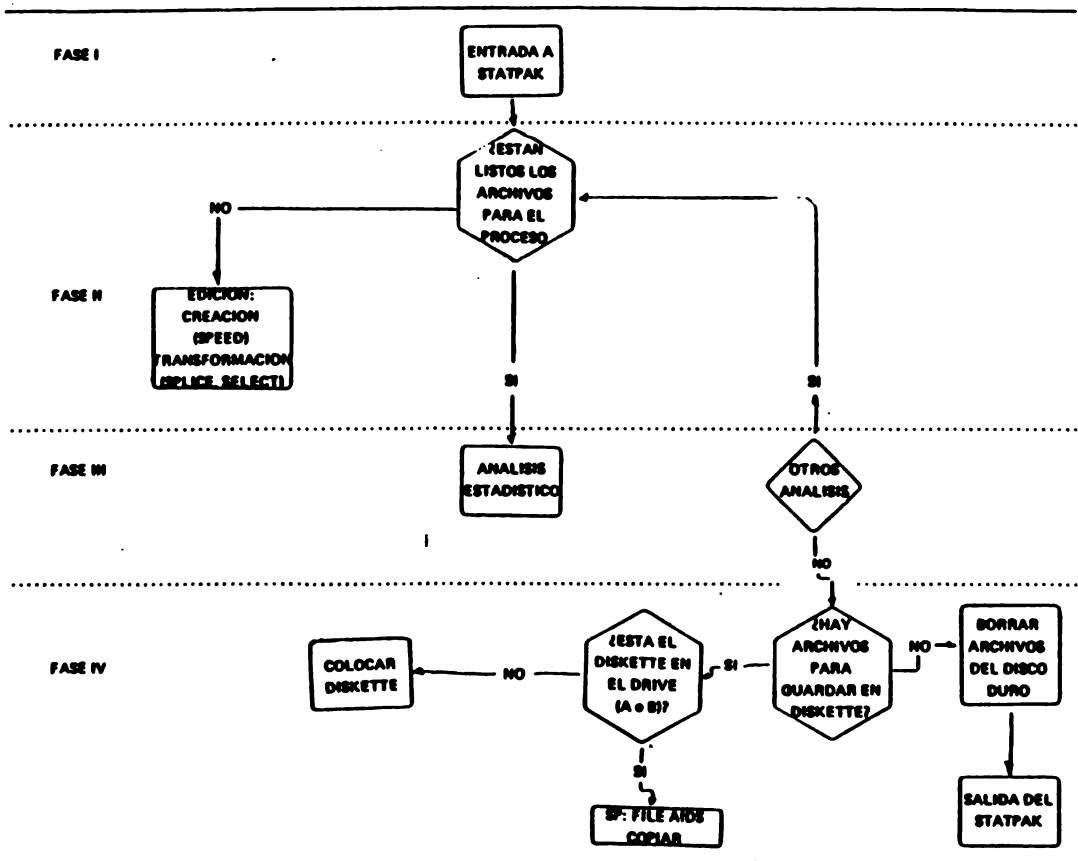
Cuadro 3.3.c

File to copy (<CR> to quit): AZUCAR.DAT
Copy to (<CR> to quit): A:AZUCAR.DAT
Copying....
File AZUCAR.DAT copied to A:AZUCAR.DAT.

File to copy (<CR> to quit):

mucho cuidado con el uso de D(elete) puesto que los errores que se cometan van a ser costosos o, por lo menos, desagradables.

En cuadro 3.4 es un flujograma tratando de mostrar el proceso lógico de una sesión de trabajo con STATPAK, obviamente en él no están contempladas "todas" las posibilidades pero puede servir para recordar algunos detalles importantes.



IV. DESCRIPCION DE PROGRAMAS

En esta sección se describirán detalladamente un conjunto de programas agrupados según la función fundamental que pueden realizar, de la siguiente manera:

- 4.1 Programas para el manejo de archivos: SPEED, SPLICE, SELECT, SORT
- 4.2 Distribuciones de probabilidad: RANUM, CONDIS1, CONDIS2, DISDIS
- 4.3 Estadística descriptiva: DSTATS, FREHIST, FREQUICK
- 4.4 Pruebas de Hipótesis que utilizan la distribución T de Student: MTEST
- 4.5 Análisis de regresión y correlación: CORREL, MLINREG, ONEVREG, POLYREG
- 4.6 Graficación: FUNGEN, QPLOT, PLOT
- 4.7 Diseños experimentales: ANOVA1, ANOVA1R, ANOVA2, ANOVA 2R1, ANOVA2R2, ANOVA3, ANOVA3R1, ANOVA3R2

La parte correspondiente a estadística no paramétrica y series de tiempo se tratará aparte, puesto que requiere no sólo el tratamiento de computación sino explicaciones estadísticas acerca de las pruebas.

4.1 Programas para el manejo de archivos

4.1.1. SPEED

Editor Ensanchado STATPAK

4.1.1.1. Descripción

Es uno de los programas editores del STATPAK y permite al usuario crear y manipular sus archivos de datos, sin tener que utilizar otros sistemas, aunque, conviene aclararlo, archivos creados en otro Software como Lotus, Wordstar, etc., pueden ser utilizados por cualquier programa SP, incluyendo el Speed.

4.1.1.2. Funciones

Speed permite las siguientes acciones:

- (1) Crear un archivo entrando los datos desde la consola
- (2) Desplegar un archivo en pantalla
- (3) Insertar, borrar, reemplazar hileras
- (4) Reemplazar caracteres
- (5) Buscar caracteres
- (6) Imprimir archivos o partes de ellos
- (7) Unir archivos horizontal o verticalmente
- (8) Dividir archivos

4.1.1.3. Limitaciones

El tamaño de los archivos que pueden crearse con Speed depende fundamentalmente de la longitud (número de caracteres) de cada registro (hilera u observación). En una microcomputadora con 128K o más de memoria RAM el producto ((longitud del registro + 4) x (número de hileras)) debe ser menor de 45.000.

De hecho el programa Speed trae la limitación de 300 hileras que puede modificarse sobre el programa directamente (se recomienda que consulte al técnico para este tipo de cambios en los programas). Así, normalmente se puede trabajar con 300 hileras y una longitud de registro máximo de 140 columnas incluyendo los espacios en blanco entre los datos.

4.1.1.4. Funcionamiento

Speed opera en dos modos diferentes:

- (1) el modo de inserción (INSERT MODE); y
- (2) el modo de comandos de edición (COMMAND MODE)

El modo de inserción sirve únicamente para entrar datos desde la consola; el modo de comandos de edición permite todas las demás tareas de edición (cambios, correcciones, etc.).

(1) El Modo de Inserción (INSERT MODE) es el modo en que opera Speed cuando se define un archivo nuevo en la declaración del archivo de entrada (Input File Declaration, sección 2.5.1). En este modo la hilera de trabajo se identifica con su número de orden y dos puntos (ejemplo: "5:" indicando que speed espera los datos correspondientes a la quinta hilera). El usuario procede a entrar los datos correspondientes a la hilera señalada y al terminar de hacerlo teclea <CR> con lo cual Speed va a la siguiente línea ("6:"). Al terminar de entrar todos los datos se da una línea en blanco (<CR> sin datos) lo cual pondrá a Speed en el modo de comandos de edición.

Nota 1: La única operación de edición permitida en el modo de inserción es el retroceso dentro de una hilera, con el fin de hacer alguna corrección, siempre y cuando no se haya dado entrada a esa hilera (no se haya tecleado <CR>). En general se recomienda entrar todos los datos (aunque vayan algunos errores) y pasar luego al modo de comandos para hacer las correcciones necesarias.

Nota 2: Aunque no se hayan cometido errores en la entrada de los datos se debe pasar al "Command Mode" para poder salir de Speed y pasar a otros programas.

(2) El Modo de Comandos de Edición (COMMAND MODE) es el modo en que se muestra Speed cuando en la declaración del archivo de entrada se define un archivo que ya existe o cuando se sale del modo de inserción. En este modo se identifica la hilera de trabajo (cuadro 4.1) con su número de orden y una flecha (ejemplo "1 →" indica que Speed espera las modificaciones que se van a hacer a la línea 1). El usuario procede a las modificaciones dando ciertos comandos (Speed Commands) que se ejecutan al teclear <CR>; uno de esos comandos es el de salida del programa Speed.

Cuando se está en el modo de comandos la pantalla muestra el archivo que se está trabajando (editando). En la línea superior aparece la llamada "línea de status" que indica el número de hileras y columnas y el nombre del archivo junto con la anotación: "'?' for help" (cuadro 4.1). Si el usuario teclea ?<CR> Speed le mostrará el resumen de comandos disponibles y una brevísima descripción de la función que realiza cada uno de ellos (cuadro 4.2). Para retornar al modo de edición se da un <CR>.

Cuadro 4.1 Modo de Comandos de Edición

```

Rows to EOF: 9   Top Row Cols: 3   File: azucar.dat
1: 9.04 15.30 12.29
2: 8.76 15.56 12.04
3: 8.08 16.71 11.21
4: 7.74 18.20 11.14
5: 7.18 17.63 10.27
6: 7.00 19.49 10.11
7: 7.90 17.25 10.97
8: 8.15 16.88 11.15
9: 7.66 16.92 10.64
-- END OF FILE --
1 ->

```

Cuadro 4.2 Resumen de Comandos Speed

```

>>-----> SPEED COMMANDS >>----->
RETURN   GO TO: next row
>        GO TO: next screen      (+ 20 Rows)
<        GO TO: previous screen (- 20 Rows)
>>       GO TO: End of File
<<       GO TO: Beginning of File
n        GO TO: row 'n'

+ [file]  INSERT: from 'file' if present, else keyboard
+-        INSERT: last row deleted: INITIAL CONTENTS OF +-
-[n]     DELETE: 'n' rows if present, else 1 row
= string  REPLACE: Current Row with 'string'
|[options] CHANGE: Add, Delete, Replace - Default: 1-99999910

#[n]     CHECK: Columns: use 'n' if present, else Current Row
@[string] FIND: 'string' - Default: -999999
P[n] [ID] PRINT: from Curr Row; 'n' rows if present, else to end
S        SAVE: all rows to Current File
S[n] file SAVE: to 'file' from Curr Row for 'n' rows, or to end
Q [opts] QUIT: SAVE to Curr File, then as per Std Exit Options
X [opts] EXIT: as above, but without file SAVE

```

9 Rows in File (Limit: 300) - Free Memory: 47K
New Row Count for '<' & '>', or <CR> to continue :

La penúltima hilera de la hoja de comandos disponibles indica cuántas hileras tiene el archivo que se está trabajando, el límite (300 en este caso) y la cantidad de memoria disponible aún para guardar datos (47 K en este ejemplo). La última se refiere a manejos de la impresora que no se requiere hacer en casos comunes.

4.1.1.5. Utilización de los comandos Speed. Con el fin de hacer el trabajo de edición en un archivo se utilizan una serie de comandos cada uno de los cuales tiene una función específica. En el cuadro 4.2 se dio una lista de ellos con una breve descripción sobre su función. En esta sección se trabajará sobre el ejemplo ya utilizado antes, AZUCAR.DAT, mostrando la acción que produce cada comando. Es conveniente, para ejercitarse, seguir este ejemplo paso por paso en la computadora.

(1) Creación del Archivo. Entre a STATPAK siguiendo las instrucciones de la sección 3.1.. Cuando la máquina pregunte por el programa a ejecutar teclee Speed <CR>, aparecerá la línea: (junto con otras)

1: .

indicando que está en modo de inserción, lista para recibir los datos de la primera hilera. Teclee los datos del cuadro 4.1. dejando un espacio entre dato y dato y <CR> al terminar cada hilera:

1:9.04 15.30 12.29 <CR>

2: .

al terminar la hilera 9 la máquina dará

10: .

teclea <CR> (línea vacía), STATPAK se pondrá en el modo de comandos de edición, mostrando en pantalla una imagen como la del cuadro 4.1..

(2) Ejercicio de edición. Siga en la computadora los siguientes pasos, observe reflexivamente el resultado de cada acción, imagine cualquier otra cosa y pruébela (para entrar cada instrucción teclee <CR>, ejemplo 5 <CR>).

Computadora C:1 + <u>.</u>	Lista para modificar la primera línea
Usuario U: <CR>	Salte a la siguiente línea
C:2 + <u>.</u>	Lista para modificar la segunda hilera
U: 5	Vaya a la quinta hilera
C:5 + <u>.</u>	Lista para modificar la quinta hilera
U: >>	Vaya al final del archivo

C:10+.

U: << Vaya al comienzo del archivo

C:1+.

U: 5

C:5+.

U: - Elimine la hilera presente (la quinta en este caso)

C:5+. Note que la hilera sexta quedó como quinta hilera

U: +- Recupere e inserte de nuevo la última hilera eliminada (quinta en este caso)

C:5+. Recuperó la quinta hilera, otra vez tiene 9

U: <<

C:1+.

U: + Pase a modo de inserción

C:1:.

U:Prueba <CR > Ponga como primera hilera del archivo la palabra prueba

C:2:. Observe la primera hilera y el número total de ellas (10)

U: <CR > Vuelva al modo de comandos

C:1+.

U: - Elimine la hilera de trabajo (primera en este caso)

C:1+. Otra vez el archivo original, mírelo

U:Ⓒ 17.63 Busque el valor 17.63 en el archivo

C:5+. Hilera donde encontró el dato

U:|17.63|***** Reemplace el valor 17.63 por 5 asteriscos

C:5+. Observe el cambio en la línea 5

Cuadro 4.3

--> PROG: SPEED FILE: AZUCAR.DAT DATE: 02-18-1986 ID: IMPRESION DE LAS 5 ULTIMAS LINEAS
PAGE 1

5: 7.18 17.63 10.27
6: 7.00 19.49 10.11
7: 7.90 17.25 10.97
8: 8.15 16.88 11.15
9: 7.66 16.92 10.64

--> PROG: SPEED FILE: AZUCAR.DAT DATE: 02-18-1986 ID: IMPRESION DE LAS LINEAS 3,4,5,6
PAGE 1

3: 8.08 16.71 11.21
4: 7.74 18.20 11.14
5: 7.18 17.63 10.27
6: 7.00 19.49 10.11

U: | ***** 10.27 | 17.63 *****

C: 5 → Observe el cambio en la línea 5

U: | ***** | 10.27

C: 5 → Vea la línea 5 de nuevo correcta

U: | 17.63 || Elimine el valor 17.63 en la línea 5

C: 5 → Observe la línea 5

U: | 7.18 | ~~17.63~~ | Adicione ~~17.63~~ luego de 7.18

C: 5 →

U: || 1986b | Adicione 1986b al comienzo de la hilera

C: 5 →

U: | 1986b || Elimine 1986b

C: 5 →

U: || b1986 | Adicione b1986 al final de la hilera

C: 5 →

U: | 1986 || Elimine 1986

C: 5 →

[Prepare la impresora]

U: P5 Impresión de las 5 últimas líneas CR
Imprima 5 líneas, a partir de la quinta, escribiendo el título dado (ver cuadro 4.3.)

C: 5 →

U: 3 Vaya a la línea 3

U: P4 Impresión de las líneas 3, 4, 5, 6.
Imprima 4 hileras a partir de la tercera.

C: 3 → Ver cuadro 4.3

U: Q Grabe el archivo y salga del programa Speed.

La computadora vuelve a la instrucción de entrada al STATPAK.

(3) Descripción de otros comandos. Además de los comandos utilizados en el ejercicio existen las siguientes posibilidades: (cuadro 4.2.)

>	Va 20 hilera adelante (inicio de la siguiente pantalla), es obvio que lo hará en archivos que tengan más de 20 observaciones.
<	Va 20 hileras hacia atrás
+ [nombre del archivo]	Inserta el archivo [nombre de archivo] debajo de la hilera en que está situado Speed
-[n]	Elimina n hileras a partir de la hilera de trabajo
#[n]	Comprueba que todas las hileras (si no se da [n]) tengan datos en todas las columnas, a partir de la hilera de trabajo. Si se da [n] hace la comprobación sólo para n hileras a partir de la de trabajo
S	Graba todas las hileras del archivo en el disco duro
S [n] [nombre]	Inserta en un archivo llamado [nombre] n hileras del archivo que está trabajando
Q	Guarda el archivo y va al programa STATPAK
Q R	Guarda el archivo y retorna a Speed
Q [nombre de programa]	Guarda el archivo y va al programa [nombre de programa]
X, X R, X [nombre de programa]	Igual que Q pero " <u>sin guardar</u> " el archivo

4.1.2. SPLICE

Editor de líneas y columnas

4.1.2.1. Descripción. Es el otro programa editor de STATPAK, como ya se anotó antes tiene con Speed dos diferencias fundamentales:

- (1) Splice no sirve para entrar datos desde la consola; y
- (2) las modificaciones que se hagan a un archivo usando Splice quedarán en otro archivo

El archivo inicial permanece sin cambios. Así, a Splice se entra con algún archivo ya grabado en disco y se sale con uno, o más, nuevos archivos, modificaciones del inicial.

4.1.2.2. Funciones. Splice permite al usuario (cuadro 4.4.) crear un nuevo archivo:

- (1) Con algunas hileras (todas las columnas) de otro (Extract Rows)
- (2) Con algunas columnas (todas las hileras) de otro (Extract Columns)
- (3) Con algunas hileras y algunas columnas de otro (Extract a Block)
- (4) Que no contenga algunas hileras de otro (Delete Rows)
- (5) Que no contenga algunas columnas de otro (Delete Columns)
- (6) Formado por la inserción de algunas hileras de un archivo en otro (Insert Rows)
- (7) Formado por la inserción de algunas columnas de un archivo en otro (Insert Columns)
- (8) Formado por el reemplazo de un bloque de datos de un archivo en otro (Replace a Block)

En todos los casos los archivos que intervienen quedan sin cambios y aparece un nuevo archivo con un nombre que el mismo usuario define.

Cuadro 4.4

----> SPLICE (StatPak Line & Column Editor) <----

- 1 - Extract Rows
- 2 - Extract Columns
- 3 - Extract a Block

- 4 - Delete Rows
- 5 - Delete Columns

- 6 - Insert Rows
- 7 - Insert Columns

- 8 - Replace a Block

- 9 - EXIT THIS PROGRAM

Enter desired Command Number ---

4.1.2.3. Limitaciones. Las mismas limitaciones de Speed. Es necesario ser cuidadoso puesto que al formar un archivo a partir de otros dos de tamaño aceptable, el nuevo puede sobrepasar las limitaciones señaladas en 4.1.1.3.

4.1.2.4. Funcionamiento. En la instrucción de entrada a STATPAK (sección 3.1.) se tecléa Splice <CR> con lo cual aparece en pantalla la información sobre posibles acciones a tomar (cuadro 4.4.), para ejecutar cualquiera de ellas se tecléa el número de opción y <CR>, para salir de Splice se da la opción 9 <CR>.

En todos los casos STATPAK pregunta por el nombre del archivo de entrada (Input File) y el de salida (Output File), es claro que el segundo se genera al operar con una opción Splice sobre el primero; además, da la posibilidad de que el usuario vea los resultados de la acción en la pantalla (Echo Output to Screen [y/n?]). Al finalizar la ejecución lo indica, mostrando además el número de hileras leídas y escritas y el comando de salida de programa (Exit Command).

4.1.2.5. Descripción de las opciones. Para mayor claridad se trabaja sobre un ejemplo, a tal efecto en el cuadro 4.5. se presentan dos archivos (utilizados anteriormente): AZUCAR.DAT y AZUCAR.RSD, que se usarán para mostrar la operación con Splice. Se da una explicación exhaustiva de las instrucciones en la opción 1 y la estrictamente necesaria para las demás, puesto que muchas cosas se repiten en todas las opciones.

- (1) Extract Rows: (cuadro 4.6) entre paréntesis cuadrado (corchetes) la indicación dada por el usuario.

Acción: Crear un archivo que contiene las tres primeras hileras de AZUCAR.DAT, ponerle el nombre WORK.DAT.

1. Enter Desired Command Number -- [1 <CR>] Se escoge la opción (1)
2. Input File (<CR> = AZUCAR.DAT):[<CR>]: Pregunta por el nombre del archivo de entrada, indicando que en caso de no indicarse ninguno tomará AZUCAR.DAT, que fue el último archivo que se trabajó. Se le responde que tome ese archivo.
3. Output File (<CR> = WORK.DAT): [<CR>]: Pregunta por el nombre para el archivo de salida, indicando que en caso de no indicarle ninguno lo nombrará WORK.DAT. Se le indica que le asigne ese nombre.
4. Echo Output to Screen (y/n)? [y <CR>]: Pregunta si se quieren ver los resultados de la opción en la pantalla a medida que se van obteniendo. Se indica que sí.
5. Begin Row (<CR> = first)? [1 <CR>]: Pregunta el número de hilera a partir de la cual comenzará la extracción,

Cuadro 4.5

AZUCAR.DAT--> PROG:SPEED FILE:AZUCAR.DAT DATE:02-02-1986 ID:

1:	9.04	15.30	12.29
2:	8.76	15.56	12.04
3:	8.08	16.71	11.21
4:	7.74	18.20	11.14
5:	7.18	17.63	10.27
6:	7.00	19.49	10.11
7:	7.90	17.25	10.97
8:	8.15	16.88	11.15
9:	7.66	16.92	10.64

--> PROG:SPEED FILE:AZUCAR.RSD DATE:02-02-1986 ID:

1:	9.04	8.797234	.2427654
2:	8.76	8.674517	8.548355E-02
3:	8.08	8.13173	-.0517292
4:	7.74	7.428464	.3115358
5:	7.18	7.697499	-.517499
6:	7	6.819598	.1804023
7:	7.9	7.876855	2.314568E-02
8:	8.15	8.051491	9.850884E-02
9:	7.66	8.032611	-.3726111

Cuadro 4.6

Enter desired Command Number -- 1

Input File (<CR>=AZUCAR.DAT):

Output File (<CR>=WORK.DAT):

Echo Output to screen (Y/N)? Y

Begin Row (<CR>=first)? 1

End Row (<CR>=last)? 3

1: 9.04 15.30 12.29

2: 8.76 15.56 12.04

3: 8.08 16.71 11.21

Done!

Input File Rows read: 4

Output File Rows writ: 3

Exit Command: <CR>=SP, R(erun), S(ystem), or PROGRAM --

Cuadro 4.7

Enter desired Command Number -- 2

Input File (<CR>=AZUCAR.DAT):

Output File (<CR>=WORK.DAT):

File exists -- re-use it (Y/N)? Y

Echo Output to screen (Y/N)? Y

Begin Column (<CR>=first)? 1

End Column (<CR>=last)? 1

1: 9.04

2: 8.76

3: 8.08

4: 7.74

5: 7.18

6: 7.00

7: 7.90

8: 8.15

9: 7.66

Done!

Input File Rows read: 9

Output File Rows writ: 9

Exit Command: <CR>=SP, R(erun), S(ystem), or PROGRAM --

indicando que de no dársele un número iniciará con la primera. Se le responde que inicie en la primera (equivale a darle simplemente <CR>)

6. End Row (<CR> = last?) [3 <CR>]: Hasta qué hilera extraer? Si no se da un número tomará la última. Se le indica que hasta la tercera
8. 1:904 15.30 12.29 Muestra las tres hileras que forman
9. 2:876 15.56 12.04 el archivo WORK.DAT, resultado de la
10. 3:808 16.71 11.21 extracción. Este archivo queda grabado en el disco duro y disponible para cualquier otro programa STATPAK
11. Done! Listo!
12. Input File Rows Read:4 Leyó 4 hileras
13. Output File Rows write:3 Escribió 3
14. Exit Command

En los próximos ejemplos al indicarse que el archivo de salida se llamará WORK.DAT, STATPAK advierte que ya existe un archivo con ese nombre y pregunta si escribe los nuevos resultados sobre él "Eliminando lo que tiene escrito actualmente" (cuidado!!). En este caso se le dirá que sí, puesto que no se requiere guardar los resultados anteriores.

(2) Extract columns: (cuadro 4.7)

Acción: Crear un archivo que solo tenga la primera columna (todas las hileras) de AZUCAR DAT.

(3) Extract a Block: (cuadro 4.8)

Acción: El nuevo archivo tendrá el bloque de información de AZUCAR.DAT formado por las hileras 2, 3, 4, 5 y las columnas 2, 3

(4) Delete Rows: (cuadro 4.9)

Acción: Crear un archivo que tenga la información de AZUCAR.DAT eliminado de él las hileras 3, 4, 5. Note la relación de esta opción con la opción (1)

(5) Delete Columns: (cuadro 4.10)

Acción: Crear un nuevo archivo a partir de AZUCAR.DAT que no tenga la segunda columna

(6) Insert Rows: (cuadro 4.11)

Cuadro 4.8

Enter desired Command Number -- 3
 Input File (<CR>=AZUCAR.DAT):
 Output File (<CR>=WORK.DAT):
 File exists -- re-use it (Y/N)? Y
 Echo Output to screen (Y/N)? Y

Begin Row (<CR>=first)? 2
 End Row (<CR>=last)? 5
 Begin Column (<CR>=first)? 2
 End Column (<CR>=last)? 3
 1: 15.56 12.04
 2: 16.71 11.21
 3: 18.20 11.14
 4: 17.63 10.27

Done!
 Input File Rows read: 6
 Output File Rows writ: 4

Exit Command: <CR>=SP, R(erun), S(ystem), or PROGRAM --

Cuadro 4.9

Enter desired Command Number -- 4
 Input File (<CR>=AZUCAR.DAT):
 Output File (<CR>=WORK.DAT):
 File exists -- re-use it (Y/N)? Y
 Echo Output to screen (Y/N)? Y

Begin Row (<CR>=first)? 3
 End Row (<CR>=last)? 5
 1: 9.04 15.30 12.29
 2: 8.76 15.56 12.04
 3: 7.00 19.49 10.11
 4: 7.90 17.25 10.97
 5: 8.15 16.88 11.15
 6: 7.66 16.92 10.64

Done!
 Input File Rows read: 9
 Rows deleted: 3
 Output File Rows writ: 6

Exit Command: <CR>=SP, R(erun), S(ystem), or PROGRAM --

Cuadro 4.10

```
Enter desired Command Number -- 5
Input File (<CR>=AZUCAR.DAT):
Output File (<CR>=WORK.DAT):
File exists -- re-use it (Y/N)? Y
Echo Output to screen (Y/N)? Y
```

```
Begin Column (<CR>=first)? 2
```

```
End Column (<CR>=last)? 2
```

```
1: 9.04 12.29
2: 8.76 12.04
3: 8.08 11.21
4: 7.74 11.14
5: 7.18 10.27
6: 7.00 10.11
7: 7.90 10.97
8: 8.15 11.15
9: 7.66 10.64
```

```
Done!
```

```
Input File Rows read: 9
```

```
Output File Rows writ: 9
```

```
Exit Command: <CR>=SP, R(erun), S(ystem), or PROGRAM --
```


Cuadro 4.11

Enter desired Command Number --- 6
 Input File (<CR>=AZUCAR.DAT):
 Output File (<CR>=WORK.DAT):
 File exists -- re-use it (Y/N)? Y
 Echo Output to screen (Y/N)? Y
 Insertion File: AZUCAR.RSD

Insert after Row (0=begin,<CR>=end)?

```

1: 9.04 15.30 11.19
2: 8.76 15.56 12.04
3: 8.08 16.71 11.21
4: 7.74 18.20 11.14
5: 7.18 17.63 10.27
6: 7.00 19.49 10.11
7: 7.90 17.25 10.97
8: 8.15 16.88 11.15
9: 7.66 16.92 10.64
10: 9.04 8.797234 .2427654
11: 8.76 8.674517 8.548355E-02
12: 8.08 8.13173 -.0517292
13: 7.74 7.428464 .3115358
14: 7.18 7.697499 -.517499
15: 7 6.819598 .1804023
16: 7.9 7.876855 2.314568E-02
17: 8.15 8.051491 9.850884E-02
18: 7.66 8.032611 -.3726111

```

Done!

Input File Rows read: 9
 Rows inserted: 9
 Output File Rows writ: 18

Exit Command: <CR>=SP, R(erun), S(ystem), or PROGRAM --

Acción: Insertar verticalmente el archivo AZUCAR.RSD (Insertion File) en AZUCAR.DAT (Input File) para crear un archivo llamado WORK.DAT (Output File). En este caso se indica que lo inserte después de la última hilera: Insert after Rows (0 = begin, CR = end)? [CR]

(7) **Insert Columns:** (cuadro 4.12)

Acción: Insertar horizontalmente el archivo AZUCAR.RSD en AZUCAR.DAT para formar WORK.DAT. En este caso se indica que lo inserte después de la última columna: Insert after Column (0 = begin, CR = end)? [CR]

(8) **Replace a Block:** (cuadro 4.13)

Acción: Insertar un bloque de información de AZUCAR.RSD en AZUCAR. DAT, el archivo resultante se denominará WORK.DAT. La computadora pregunta en qué hilera y columna de AZUCAR.DAT (Input File) inicia el bloque a reemplazar y asume que va hasta el final del archivo. En el ejemplo se le indica que inicia en hilera 6, columna 2, así que el bloque a reemplazar es:

	2	3
6	(6.1)	(6.3)
7	(7.2)	(7.3)
8	(8.2)	(8.3)
9	(9.2)	(9.3)

Formado por cuatro hileras y dos columnas

Luego toma "las primeras" cuatro hileras y dos columnas de AZUCAR.RSD para reemplazar el bloque. Observe cuidadosamente los dos archivos originales y el resultante

Cuadro 4.12

```

Enter desired Command Number -- 7
Input File (<CR>=AZUCAR.DAT):
Output File (<CR>=WORK.DAT):
File exists -- re-use it (Y/N)? Y
Echo Output to screen (Y/N)? Y
Insertion File: AZUCAR.RSD

```

```

Insert after Column (0=begin,<CR>=end)?

```

```

 1: 9.04 15.30 12.29 9.04 8.797234 .2427654
 2: 8.76 15.56 12.04 8.76 8.674517 8.548355E-02
 3: 8.08 16.71 11.21 8.08 8.13173 -.0517292
 4: 7.74 18.20 11.14 7.74 7.428464 .3115358
 5: 7.18 17.63 10.27 7.18 7.697499 -.517499
 6: 7.00 19.49 10.11 7 6.819598 .1804023
 7: 7.90 17.25 10.97 7.9 7.876855 2.314568E-02
 8: 8.15 16.88 11.15 8.15 8.051491 9.850884E-02
 9: 7.66 16.92 10.64 7.66 8.032611 -.3726111

```

```

Done!

```

```

Input File Rows read: 9
Total Insert File Rows: 9
Output File Rows writ: 9

```

```

Exit Command: <CR>=SP, R(erun), S(ystem), or PROGRAM --

```

Cuadro 4.13

Enter desired Command Number --- 8

Input File (<CR>=AZUCAR.DAT):
 Output File (<CR>=WORK.DAT):
 File exists -- re-use it (Y/N)? Y
 Echo Output to screen (Y/N)? Y
 Replacement File: AZUCAR.RSD

Begin Row (<CR>=first)? 6
 Begin Column (<CR>=first)? 2

1:	9.04	15.30	12.29
2:	8.76	15.56	12.04
3:	8.08	16.71	11.21
4:	7.74	18.20	11.14
5:	7.18	17.63	10.27
6:	7.00	9.04	8.797234
7:	7.90	8.76	8.674517
8:	8.15	8.08	8.13173
9:	7.66	7.74	7.428464

Done!

Input File Rows read: 9
 Replacement File Rows used: 4
 Output File Rows writ: 9

Exit Command: <CR>=SP, R(erun), S(ystem), or PROGRAM --

4.1.3. SELECT

Programa para transformación de datos y/o selección de observaciones

4.1.3.1. Descripción. Select realiza operaciones matemáticas sobre las variables y/o entre ellas y permite además seleccionar subconjuntos de datos que cumplan con hasta 10 requisitos simultáneamente. Los resultados generan un nuevo archivo, quedando intacto el original.

4.1.3.2. Funciones. Con la utilización de Select es posible ejecutar las siguientes acciones:

- (1) Transformación de variables
- (2) Creación de nuevas variables a partir de las originales
- (3) Selección observaciones (hileras) que satisfacen determinados requisitos respecto a los valores que toman en ellas las variables (columnas) de interés
- (4) Conservación, para uso posterior, de un archivo conteniendo instrucciones para selección o transformaciones (de utilidad cuando se trata de transformaciones o selecciones utilizadas muy a menudo)

4.1.3.3. Descripción de las funciones

- (1) (2) Transformación y/o creación de nuevas variables. Estas dos acciones se realizan mediante el uso de expresiones matemáticas; estas expresiones pueden estar formadas por uno o más de los siguientes componentes, separados por espacios en blanco.
 - a) Variables: identificadas con el número de campo asignado a cada una de ellas entre paréntesis.
 - b) Operadores suma (+), resta (-), multiplicación (x), división (/), división entera(\) y exponenciación (^)
 - c) Funciones matemáticas: escritas con una palabra clave y el número de identificación de la variable entre paréntesis

Palabra clave	Función matemática
LOG (X)	log (x)
ARCTAN (X)	Arcotangente (x)
EXP (X)	
ABS (X)	x
SQR (X)	\sqrt{x} (x+1/2)
SIN (X)	Seno (x)
COS (X)	Coseno (x)
TAN (X)	Tangente (x)

X = número de identificación de la variable
x = variable x

d) Constantes: números que participan en las transformaciones definidas

- (3) Selección de observaciones. Select permite realizar operaciones lógicas de comparación (pruebas de selección) con el fin de seleccionar determinadas observaciones de un archivo. Las operaciones lógicas permitidas son > , >= , < , <= y pueden ejecutarse tanto sobre variables alfabéticas como sobre numéricas. Es posible hacer hasta 10 pruebas de selección en un solo paso y el resultado es un archivo cuyas observaciones satisfacen "simultáneamente" (operación lógica "Y") las condiciones de las pruebas. Cada prueba de selección se define dando: 1) el número de columnas (variables) sobre la cual se decidirá; 2) información de si la variable en cuestión es alfabética o numérica; 3) límites mínimo y máximo admitidos; y 4) la orden de guardar (selección positiva) o eliminar (selección negativa) las observaciones que

satisfacen los límites (se recomienda estudiar atentamente el ejemplo al final de la sección).

- (4) Archivo de definición. Es una opción que da STATPAK para guardar las instrucciones de selección y/o transformación usadas cuando se van a repetir muchas veces, sobre diferentes conjuntos de datos. Es claro que para utilizar el mismo archivo de definición se requiere operar sobre archivos de datos con idéntica estructura.

4.1.3.4. Funcionamiento. La operación de Select ofrece muchas alternativas, en general se puede listar un orden de proceso, tomando en cuenta la información que el programa va requiriendo, de la siguiente forma:

- (1) Teclee Select <CR> en la primera instrucción STATPAK
- (2) Si ya se tiene un archivo con la definición de las transformaciones y/o selecciones del nombre; sino teclee <CR>
- (3) Indique el número de columnas del archivo de entrada que van al archivo de salida como entraron o a través de transformaciones (m)
- (4) Input File Declaration y Field Assignment (sección 2.5.)
- (5) Nombre del archivo de salida
- (6) Número de columnas en el archivo de salida (P). P puede ser menor, igual o mayor que m
- (7) Cuántas pruebas de selección se precisan
- (8) Descripción de cada una de las pruebas de selección
 - (a) Campo (variable) sobre el cual se hace la selección: se debe indicar el número de la columna "en la asignación de campos"
 - (b) Campo numérico o alfabético (N, A): indicar el tipo de valores que tome el campo indicado en (a)

- (c) Valor de los límites inferior y superior para la selección. Las pruebas incluirán los límites a menos que explícitamente se indique otra cosa. La selección de hecho consiste en guardar o eliminar las observaciones que tienen valores, en la columna definida, dentro del intervalo acá indicado.
- (d) Indicar si la selección es positiva o negativa (P, N). En el primer caso se guardan las observaciones que satisfacen los límites, en el segundo se eliminan.
- (9) Si se quiere corregir algo en las selecciones descritas en el punto 8 STATPAK da en este momento la posibilidad de hacerlo. Así, si se comete algún error en el paso 8 se puede seguir adelante y corregir en este paso 9
- (10) Descripción de las transformaciones. Cada transformación se indica escribiendo la expresión matemática que dará origen a cada columna en el archivo de salida. Las columnas que no sufren modificaciones se deben indicar también explícitamente. El programa pregunta por la definición de las columnas de salida (Output Column Definitions). En este momento se escriben las expresiones matemáticas deseadas
- (11) Cuando en la expresión matemática aparecen funciones trigonométricas, STATPAK pregunta si los ángulos están medidos en grados, radianes o grades. Si aparece la función logarítmica, pregunta por la base. La primera definición que se dé, será tomada en las siguientes transformaciones.
- (12) Posibilidad de corregir las transformaciones, igual que en el caso de las pruebas de selección.
- (13) Finalmente Select pregunta al usuario si quiere guardar las instrucciones de selección y transformación en un

archivo de definición (ver (1)) para posterior uso.

4.1.3.5. Ejemplos

Se utiliza un conjunto de datos presentados en Steel and Torrie: "Principles and Procedures of Statistics", pág. 282. Los datos corresponden a un estudio sobre el efecto del contenido de minerales, en la hoja de tabaco, sobre su facilidad de quemado medida en tiempo de quemado. El cuadro 4.14 muestra el archivo listo para ser procesado.

Ejemplo (1), Transformaciones: como primer ejemplo se trabaja con el cálculo de valores esperados y residuales a partir de una regresión múltiple. Se pretende mostrar algunas de las posibilidades que ofrece Select. El cuadro 4.15 muestra los resultados de la regresión, los cuales indican que la ecuación estimada es:

$$Y = 2.0119 - .5612 x_1 - .3481 x_2 + .1780 x_3$$

En el cuadro 4.16 se listan las instrucciones necesarias para hacer el trabajo. Con flechas se señalan los siguientes puntos de interés:

- (1) Entran cuatro variables, salen seis (se crean dos)
- (2) El nombre del archivo de salida es TABACO.TRI. En el cuadro 4.17 se presenta el archivo TABACO.TRI, resultado de las transformaciones.
- (3) No se piden pruebas de selección (se da < CR >)
- (4) Las cuatro variables originales quedan igual en el nuevo archivo.
- (5) las variables 5 y 6 se generan con expresiones matemáticas. La primera tiene los valores predichos (Y), la segunda los residuales (e). Observe el cambio de signos y compare las ecuaciones con los resultados del cuadro 4.15
- (6) Instrucción para corrección de errores, si se quiere corregir alguna columna se da su número y <CR> si todo está bien se da solamente <CR> como en este ejemplo

Cuadro 4.14

PROG: SHEFD FILE: TABACO.DAT DATE: 02-26-1985 ID: PAGE 1

1:	3.05	1.35	5.67	.34
2:	4.22	1.35	4.86	.11
3:	3.34	.26	4.19	.33
4:	3.77	0.23	4.42	.68
5:	3.52	1.10	3.17	.13
6:	3.54	0.76	2.76	.00
7:	3.74	1.59	3.81	.03
8:	3.78	.39	3.23	.11
9:	2.92	.39	3.41	1.53
10:	3.10	0.61	6.16	.77
11:	2.86	0.81	5.40	1.17
12:	2.78	.61	1.62	1.01
13:	2.22	.85	4.89	.89
14:	2.67	.9	3.59	1.4
15:	3.12	.92	5.86	1.05
16:	3.03	.97	6.6	1.15
17:	2.45	.11	0.51	1.49
18:	4.12	.62	5.31	.51
19:	4.61	.51	3.15	.18
20:	3.54	.45	0.45	.34
21:	4.12	1.75	6.17	.36
22:	2.93	.25	3.38	.89
23:	2.66	.31	3.51	.91
24:	3.17	.20	3.08	.92
25:	2.79	.24	3.72	1.33
26:	2.51	.20	3.64	1.33
27:	3.74	2.27	6.5	.23
28:	3.13	1.48	4.38	1.26
29:	3.49	.25	4.71	.73
30:	2.94	2.22	4.58	.23

Cuadro 4.15

--> PROG:MLINREG FILE:TABACU.DAT DATE:02-26-1986 ID:ANALISIS DE REGRESION
PAGE 1


TERM	COEFFICIENT	STD. ERROR	T-STATISTIC	PART. CORR	CONTR. R-SQ
B 0	2.011922	.3456057	5.821438	---	---
B 1	-.5612166	8.602672E-02	-6.523749	0.6208	0.4357
B 2	-.3480997	9.030538E-02	-3.854695	0.3637	0.1521
B 3	.1780537	5.024873E-02	3.543446	0.3257	0.1286


	SUM SQ	DEG FR	MEAN SQ
DUE TO REGRESSION	4.992938	3	1.664313
ABOUT REGRESSION	1.811248	26	.0696634
TOTAL	6.804186	29	.2346271


R-SQUARED: .7338038 CORRECTED R-SQUARED: .7030889
F-TEST: 23.89078 STD ERROR OF REG: .2639383


Cuadro 4.16

----> S E L E C T <----

Definition File Name (<CR> if none): 
 Number of Input Fields (1 - 120)? 4

Input File (Q=Quit, <CR>=TABACO.DAT):
 4 Data Field(s)
 4 File Columns in TABACO.DAT
 Field Assignment(s) (<CR>=1:1): 

Output File Name ('Q' to Quit): TABACO.TR1
 Number of Output Columns (1 - 120)? 6 

Number of Selection Tests (0 - 10)? 

----> OUTPUT COLUMN DEFINITIONS <----

Column 1 (of 6) Definition (<CR>= (1)):

=

Column 2 (of 6) Definition (<CR>= (2)):

=


Column 3 (of 6) Definition (<CR>= (3)):

=


Column 4 (of 6) Definition (<CR>= (4)):

=

Column 5 (of 6) Definition (<CR>= (4)):

= $2.0119 - .5612 * (1) - .3481 * (2) + .1780 * (3)$ 

Column 6 (of 6) Definition (<CR>= (4)):

= $(4) - 2.0119 + .5612 * (1) + .3481 * (2) - .1780 * (3)$ 

Definition to change (<CR> to continue)? 

Save Definition data in a file (Y/N)? N

WORKING *****

Input Records read: 30

Output Records writ: 30

Exit Command: <CR>=SP, R(erun), S(ystem), or PROGRAM --

Cuadro 4.17

--> PROG: SPEED FILE: TABACO.TRI DATE: 02-26-1986 ID: PAGE 1

1:	3.05	1.45	5.67	.34	34.52449	-25.9988
2:	4.22	1.35	4.86	.11	38.74733	-35.70172
3:	3.34	.26	4.19	.38	5.64511	-4.262444
4:	3.77	0.23	4.42	.68	5.992808	-3.586652
5:	3.52	1.10	3.17	.18	17.15667	-14.94731
6:	3.54	0.76	2.76	.00	10.53328	-10.53328
7:	3.74	1.59	3.81	.08	31.43727	-29.62475
8:	3.78	.39	3.23	.11	7.044186	-6.520403
9:	2.92	.39	5.44	1.53	9.216981	.2614783
10:	3.10	0.64	6.16	.77	17.45377	-8.043263
11:	2.96	0.82	5.48	1.17	18.05917	-3.018688
12:	2.78	.64	4.62	1.01	11.71771	-3.415609
13:	2.22	.85	4.49	.89	11.76194	-4.221299
14:	2.67	.9	5.59	1.4	18.73065	.0752303
15:	3.12	.92	5.86	1.05	23.56797	-5.906395
16:	3.03	.97	6.6	1.15	27.08703	-4.779259
17:	2.45	.18	4.51	1.49	3.405505	-.4420281
18:	4.12	.62	5.31	.51	19.47626	-12.55869
19:	4.61	.51	5.16	.18	17.60184	-15.41814
20:	3.94	.45	4.45	.34	11.54083	-8.858285
21:	4.12	1.79	6.17	.36	63.26423	-46.88333
22:	2.93	.25	3.33	.89	3.979211	-1.695705
23:	2.66	.31	3.51	.91	4.44484	-1.810985
24:	3.17	.20	3.08	.92	3.166621	-1.370119
25:	2.79	.24	3.98	1.35	4.242062	-.644301
26:	2.61	.20	3.64	1.33	3.150949	-.6238426
27:	3.74	2.27	6.5	.23	76.07578	-63.38352
28:	3.13	1.48	4.28	1.26	27.31939	-2.337781
29:	3.49	.25	4.71	.73	6.390107	-3.390191
30:	2.94	2.22	4.58	.23	40.64131	-33.76597

Ejemplo (2), Una Prueba de Selección:
(cuadros 4.18 y 4.20). En este ejemplo es interesante observar lo siguiente:

- (1) Entran y salen cuatro variables
- (2) El archivo de entrada es TABACO.SRT (los mismos datos de TABACO.DAT pero ordenados según la primera columna en orden ascendente, cuadro 4.18). El de salida TABACO.TRY
- (3) Se hará una prueba de selección.
- (4) Definición de la selección
- (5) Descripción de la selección: STATPAK indica que se guardarán (selección positiva) las observaciones cuya primera variable tome un valor mayor o igual a 2.93 y menor o igual a 3.74, luego pregunta si se quiere modificar alguno de los parámetros de la definición
- (6) Se indica que no hay transformaciones, todas las columnas quedan con los valores que tenían.

El archivo resultante (cuadro 4.20) contiene 12 observaciones que cumplieron con las condiciones puestas a la primera variable. Compare este archivo con TABACO.SRT detenidamente para comprender el trabajo realizado.

Ejemplo (3), Dos Pruebas de Selección: en este ejemplo se dan restricciones sobre dos columnas, la uno y la cuatro. Las flechas en el cuadro 4.21 indican puntos importantes:






- (1) Entran y salen cuatro variables
- (2) El archivo de entrada es TABACO.SRT (Cuadro 4.18), el de salida TABACO.TR3
- (3) Se hace la selección con dos restricciones
- (4) Definición de las selecciones, la primera es igual a la del ejemplo 2
- (5) No se da límite inferior, STATPAK toma -1×10^{-37} (si no se da superior toma 1×10^{37})

Cuadro 4.18


--> PROG: SPEED FILE: TABACO.SRT DATE: 02-26-1986 ID: PAGE 1

1:	2.22	.85	4.49	.89
2:	2.45	.18	4.51	1.49
3:	2.61	.20	3.64	1.33
4:	2.66	.31	3.51	.91
5:	2.67	.9	5.59	1.4
6:	2.78	.64	4.62	1.01
7:	2.79	.24	3.98	1.35
8:	2.86	0.82	5.48	1.17
9:	2.92	.39	5.44	1.53
10:	2.93	.25	3.38	.89
11:	2.94	2.22	4.58	.23
12:	3.03	.97	6.6	1.15
13:	3.05	1.45	5.67	.34
14:	3.10	0.64	6.16	.77
15:	3.12	.92	5.86	1.05
16:	3.13	1.48	4.28	1.26
17:	3.17	.20	3.08	.92
18:	3.34	.26	4.19	.38
19:	3.49	.25	4.71	.73
20:	3.52	1.10	3.17	.18
21:	3.54	0.76	2.76	.00
22:	3.74	2.27	6.5	.23
23:	3.74	1.59	3.81	.08
24:	3.77	0.23	4.42	.68
25:	3.78	.39	3.23	.11
26:	3.94	.45	4.45	.34
27:	4.12	.62	5.31	.51
28:	4.12	1.79	6.17	.36
29:	4.22	1.35	4.86	.11
30:	4.61	.51	5.16	.18


Cuadro 4.19

Definition File Name (<CR> if none): 
 Number of Input Fields (1 - 120)? 4 
 Input File (Q=Quit, <CR>=TABACO.SRT): 
 4 Data Field(s)
 4 File Columns in TABACO.SRT
 Field Assignment(s) (<CR>=1:1):
 Output File Name ('Q' to Quit): TABACO.TR2
 Number of Output Columns (1 - 120)? 4 
 Number of Selection Tests (0 - 10)? 1 


For non-inclusive Lower Limit, prefix with '>'
 For non-inclusive Upper Limit, prefix with '<'

--- SELECTION # 1 ---
 Field # (<CR>=0)? 1 
 Numeric or Alpha (N,A)? N,
 Lower Limit (<CR>=none): 2.93
 Upper Limit (<CR>=none): 3.74
 Positive or Negative selection (P,N)? P

----> SELECTION RECAP <----

 1 - Field: 1(Num) Lower: 2.93 Upper: 3.74 Fos
 Selection to change (<CR> to continue)?

----> OUTPUT COLUMN DEFINITIONS <----

Column 1 (of 4) Definition (<CR>= (1)):
 =
 Column 2 (of 4) Definition (<CR>= (2)):
 =
 Column 3 (of 4) Definition (<CR>= (3)): 
 =
 Column 4 (of 4) Definition (<CR>= (4)):
 =
 Definition to change (<CR> to continue)?
 Save Definition data in a file (Y/N)? N
 WORKING*****.....

Input Records read: 30
 Output Records writ: 12
 Exit Command: <CR>=SP, R(erun), S(ystem), or PROGRAM --- SPEED 0

Cuadro 4.20

--> PROG: SPEED FILE: TABACO.TR2 DATE: 02-25-1986 ID: PAGE 1

1:	2.93	.25	3.38	.89
2:	2.94	2.22	4.58	.23
3:	3.03	.97	6.6	1.15
4:	3.05	1.45	5.67	.34
5:	3.10	0.64	6.16	.77
6:	3.12	.92	5.86	1.05
7:	3.13	1.48	4.28	1.26
8:	3.17	.20	3.08	.92
9:	3.34	.26	4.19	.38
10:	3.49	.25	4.71	.73
11:	3.52	1.10	3.17	.18
12:	3.54	0.76	2.76	.00


- (6) Se pide que el límite superior no incluya el valor .92
- (7) Descripción de las selecciones: se guardarán las observaciones que simultáneamente cumplan con que la primera variable toma valores mayores o iguales a 2.93 y menores o iguales a 3.74 y la cuarta toma valores estrictamente menores que .92
- (8) No se transforman los valores de las variables. El archivo de salida, cuadro 4.22, tiene 8 observaciones, si se le compara con el anterior, cuadro 4.20, se observará que no guardó las observaciones números 3, 6, 7, 8 que tienen valores mayores o iguales a .92 en la 4a. columna.


Ejemplo (4), Pruebas de Selección Negativas:


con base en el ejemplo (3) se muestra el resultado de hacer negativa la selección (eliminar las observaciones que satisfacen). En el cuadro 4.23 se dan las instrucciones, en el 4.24 los resultados. Observe, señalado con flechas:


- (1) Entra el archivo TABACO.SRT, mismo del ejemplo (3) se genera TABACO.TR4
- (2) Las pruebas de selección son idénticas a las del ejemplo (3) (cuadro 4.21) excepto por la elección N en la segunda selección
- (3) Observe la palabra Neg en la descripción de las selecciones. De la comparación de los cuadros 4.20, 4.22 y 4.24 se puede observar que mientras la selección positiva dejó las observaciones 1, 2, 4, 5, 9, 10, 11 y 12, cuyos valores en la última columna son estrictamente menores que .92 (TABACO.TR3), la selección negativa las eliminó (TABACO.TR4) dejando las observaciones números 3, 6, 7, 8 que no cumplen la condición

Cuadro 4.21


Definition File Name (<CR> if none): 
 Number of Input Fields (1 - 120)? 4




 Input File (Q=Quit, <CR>=TABACO.TR2): TABACO.SRT
 4 Data Field(s)
 4 File Columns in TABACO.SRT
 Field Assignment(s) (<CR>=1:1): 

 Output File Name ('Q' to Quit): TABACO.TR3
 Number of Output Columns (1 - 120)? 4 


 Number of Selection Tests (0 - 10)? 2 

For non-inclusive Lower Limit, prefix with '>'
 For non-inclusive Upper Limit, prefix with '<'

--- SELECTION # 1 --- 
 Field # (<CR>=0)? 1
 Numeric or Alpha (N,A)? N
 Lower Limit (<CR>=none): 2.93
 Upper Limit (<CR>=none): 3.74
 Positive or Negative selection (P,N)? P

--- SELECTION # 2 --- 
 Field # (<CR>=0)? 4
 Numeric or Alpha (N,A)? N
 Lower Limit (<CR>=none): 
 Upper Limit (<CR>=none): <.92 
 Positive or Negative selection (P,N)? P

---> SELECTION RECAP <---

 1 - Field: 1(Num) Lower: 2.93
 2 - Field: 4(Num) Lower: -1E+37
 Selection to change (<CR> to continue)?

Upper: 3.74 Pos
 Upper: <.92 Pos

Cuadro 4.21 (Continuación)

```

      ---> OUTPUT COLUMN DEFINITIONS <---
Column 1 ( of 4 ) Definition (<CR>= (1) ):
=
Column 2 ( of 4 ) Definition (<CR>= (2) ):
=
Column 3 ( of 4 ) Definition (<CR>= (3) ):
=
Column 4 ( of 4 ) Definition (<CR>= (4) ):
=
Definition to change (<CR> to continue)?
Save Definition data in a file (Y/N)? N
WORKING .....**.**...***.....

Input Records read: 30
Output Records writ: 8
Exit Command: <CR>=SP, R(erun), S(ystem), or PROGRAM --

```

Cuadro 4.22

```

--> PROG:SPEED FILE:TABACO.TR3 DATE:02-25-1986 ID:PAGE 1

```


```


1: 2.93 .25 3.38 .89
2: 2.94 2.22 4.58 .23
3: 3.05 1.45 5.67 .34
4: 3.10 0.64 6.16 .77
5: 3.34 .26 4.19 .38
6: 3.49 .25 4.71 .73
7: 3.52 1.10 3.17 .18
8: 3.54 0.76 2.76 .00

```

Cuadro 4.23

Definition File Name (<CR> if none):
 Number of Input Fields (1 - 120)? 4


Input File (Q=Quit, <CR>=TABACO.TR2): TABACO.SRT 
 4 Data Field(s)
 4 File Columns in TABACO.SRT
 Field Assignment(s) (<CR>=1:1):

Output File Name ('Q' to Quit): TABACO.TR4 
 Number of Output Columns (1 - 120)? 4

Number of Selection Tests (0 - 10)? 2

For non-inclusive Lower Limit, prefix with '>'
 For non-inclusive Upper Limit, prefix with '<'

--- SELECTION # 1 ---
 Field # (<CR>=0)? 1
 Numeric or Alpha (N,A)? N
 Lower Limit (<CR>=none): 2.93
 Upper Limit (<CR>=none): 3.74
 Positive or Negative selection (P,N)? P


--- SELECTION # 2 ---
 Field # (<CR>=0)? 4
 Numeric or Alpha (N,A)? N
 Lower Limit (<CR>=none):
 Upper Limit (<CR>=none): <.92
 Positive or Negative selection (P,N)? N 

Cuadro 4.23 (Continuación)

```

,---> SELECTION RECAP <---
  1 - Field:  1( Num )   Lower: 2.93
  2 - Field:  4( Num )   Lower: -1E+37
Selection to change <CR> to continue)?
Upper: 3.74 Pos
Upper: <.92 Neg

```



```

----> OUTPUT COLUMN DEFINITIONS <----

```

```

Column 1 ( of 4 ) Definition <CR>= (1) ):
=
Column 2 ( of 4 ) Definition <CR>= (2) ):
=
Column 3 ( of 4 ) Definition <CR>= (3) ):
=
Column 4 ( of 4 ) Definition <CR>= (4) ):
=
Definition to change <CR> to continue)?
Save Definition data in a file (Y/N)? N
WORKING .....*...***.....

```

```

Input Records read:  30
Output Records writ:  4
Exit Command: <CR>=SP, R(erun), S(ystem), or PROGRAM -- SPEED C

```

Cuadro 4.24

```

--> PROG:SPEED  FILE:TABACO.TR4  DATE:02-26-1986  ID: PAGE  1

```

```

1:  3.03  .97  6.6  1.15
2:  3.12  .92  5.86  1.05
3:  3.13  1.48  4.28  1.26
4:  3.17  .20  3.08  .92

```

4.1.4. SORT

Programa Ordenador de Archivos

4.1.4.1. Descripción: A partir de un archivo original SORT genera uno nuevo que contiene la misma información, pero cuyas observaciones (hileras) están escritas en otro orden. El orden del nuevo archivo depende de los valores que toma una de sus variables, definida por el usuario y llamada el campo clave (KEY FIELD), los cuales son escritos en orden ascendente o descendente.

4.1.4.2. Funciones: SORT permite preparar archivos para procesos como graficación o para mejorar su presentación. Ordena los valores (alfabéticos o numéricos) de la variable clave y los escribe junto con los demás datos que formaban la observación a la que pertenece cada valor. La variable clave es la primera columna del nuevo archivo.

4.1.4.3. Funcionamiento: Los pasos en la operación de SORT son los siguientes:

- (1) Teclee SORT <CR> en la primera instrucción STATPAK, la computadora pregunta por el número de columnas en el archivo de salida advirtiéndole que en él la variable clave va como primera columna
- (2) Entre el número de columnas en el archivo de salida
- (3) Definición del archivo de entrada (INPUT FILE DECLARATION)
- (4) Asignación de campos (FIELD ASSIGNMENT), en la asignación se da la variable (columna) que se desea tomar como clave en el primer lugar
- (5) Indique si la variable clave toma valores alfabéticos o numéricos (A/N)
- (6) Indique si el ordenamiento de la variable clave se hará en orden ascendente o descendente
- (7) Defina el nombre del archivo de salida

4.1.4.4. Ejemplos: Sobre los mismos datos usados en la ejemplificación de SELECT, TABACO.DAT se corrieron los siguientes ejemplos:

Ejemplo (1): Ordenamiento en orden ascendente con base en la primera columna. Observe las siguientes anotaciones, indicadas con flechas sobre el cuadro (4.25)

- (1) El archivo de entrada es TABACO.DAT, el de salida TABACO SRT
- (2) La asignación es 1:1, esto indica que SORT operará tomando como variable clave para el ordenamiento la primera columna de TABACO.DAT.
- (3) Puesto que la variable clave toma valores numéricos el SORT es de este tipo
- (4) Se le indica que ordene de menor a mayor
- (5) Avisa que terminó y cuántos registros (Observaciones) quedaron en el archivo de salida

Compare los cuadros 4.14 y 4.26 para que tenga total comprensión del trabajo de SORT. Especialmente observe que un valor de la variable clave arrastra el resto de valores de su observación, así no hay ningún tipo de confusión en la información.

Ejemplo (2): Ordenamiento en orden descendente con base en la cuarta columna, cuadro 4.27:

- (1) Archivo de entrada TABACO.DAT, de salida TABACO.SRT. Puesto que en la elaboración de los ejemplos se usó ese mismo nombre en el ejemplo anterior STATPAK advierte que TABACO.SRT ya existe y pregunta si escribe encima de él el nuevo archivo ordenado (obviamente perdiéndose la información que tiene en este momento)
- (2) Se asignan las variables en el orden 4, 1, 2, 3, esto implica que la variable clave ahora es la cuarta y que en el archivo de salida figure como primera. Tenga cuidado con este cambio en el orden de las variables cuando utilice en otros análisis archivos de salida de SORT
- (3) Se indica hacer ordenamiento de mayor a menor de la variable clave

El archivo resultante se presenta en el cuadro 4.28.

Cuadro 4.25

of Columns in Output File? 4

Data Column #1 is Sort Key

Input File (Q=Quit, <CR>=TABACO.DAT):

4 Data Field(s)

4 File Columns in TABACO.DAT

Field Assignment (s) (<CR>=1:1):

Numeric or Alpha Sort Key (N,A)? N

Ascending or Descending (A,D)? A

Output File Name ('Q' to Quit): TABACO.SRT

Determining File Size.....

30 records to sort.

Reading Data.....

Sorting.....

Writing.....

Done - 30 Records writ to Output.

Exit Command: <CR>=SP, R(erun), S(ystem), or PRUGRAM --

Cuadro 4.26

--> PROG: SPEED FILE: TABACO.SRT DATE: 02-26-1986 ID: PAGE 1

1:	2.22	.85	4.49	.89
2:	2.45	.18	4.51	1.49
3:	2.61	.20	3.64	1.33
4:	2.66	.31	3.51	.91
5:	2.67	.9	5.59	1.4
6:	2.78	.64	4.62	1.01
7:	2.79	.24	3.98	1.35
8:	2.86	0.82	5.48	1.17
9:	2.92	.39	5.44	1.53
10:	2.93	.25	3.38	.89
11:	2.94	2.22	4.58	.23
12:	3.03	.97	6.6	1.15
13:	3.05	1.45	5.67	.34
14:	3.10	0.64	6.16	.77
15:	3.12	.92	5.86	1.05
16:	3.13	1.48	4.28	1.26
17:	3.17	.20	3.08	.92
18:	3.34	.26	4.19	.38
19:	3.49	.25	4.71	.73
20:	3.52	1.10	3.17	.18
21:	3.54	0.76	2.76	.00
22:	3.74	2.27	6.5	.23
23:	3.74	1.59	3.81	.08
24:	3.77	0.23	4.42	.68
25:	3.78	.39	3.23	.11
26:	3.94	.45	4.45	.34
27:	4.12	.62	5.31	.51
28:	4.12	1.79	6.17	.36
29:	4.22	1.35	4.86	.11
30:	4.61	.51	5.16	.18

Cuadro 4 27

```

# of Columns in Output File? 4
Data Column #1 is Sort Key
Input File (Q=Quit, <CR>=TABACO.1R4): TABACO.DAT
  4 Data Field(s)
  4 File Columns in TABACO.DAT
Field Assignment(s) (<CR>=1:1): 4 1 2 3
Numeric or Alpha Sort Key (N,A)? N
Ascending or Descending (A,D)? D
Output File Name ('Q' to Quit): TABACO.SRT
TABACO.SRT exists: re-use it (Y,N)? Y

Determining File Size.....
  30 records to sort.

Reading Data.....
Sorting.....
Writing.....

Done - 30 Records writ to Output.
Exit Command: <CR>=SP, R(erun), S(ystem), or PROGRAM --

```

Cuadro 4.28

--> PROG: SPEED FILE: TABACO.SRT DATE: 02-26-1986 ID: PAGE 1

1:	1.53	2.92	.39	5.44
2:	1.49	2.45	.18	4.51
3:	1.4	2.67	.9	5.59
4:	1.35	2.79	.24	3.98
5:	1.33	2.61	.20	3.64
6:	1.26	3.13	1.48	4.28
7:	1.17	2.86	0.82	5.48
8:	1.15	3.03	.97	6.6
9:	1.05	3.12	.92	5.86
10:	1.01	2.78	.64	4.62
11:	.92	3.17	.20	3.08
12:	.91	2.66	.31	3.51
13:	.89	2.93	.25	3.38
14:	.89	2.22	.85	4.49
15:	.77	3.10	0.64	6.16
16:	.73	3.49	.25	4.71
17:	.68	3.77	0.23	4.42
18:	.51	4.12	.62	5.31
19:	.38	3.34	.26	4.19
20:	.36	4.12	1.79	6.17
21:	.34	3.74	.45	4.45
22:	.34	3.05	1.45	5.67
23:	.23	2.94	2.22	4.58
24:	.23	3.74	2.27	6.5
25:	.18	4.61	.51	5.16
26:	.18	3.52	1.10	3.17
27:	.11	3.78	.39	3.23
28:	.11	4.22	1.35	4.86
29:	.08	3.74	1.59	3.81
30:	.00	3.54	0.76	2.76

4.2 Programas para generación de número aleatorios y manejo de distribuciones probabilísticas

4.2.1. RANUM

Generador de Número Aleatorios

4.2.1.1. Descripción: RANUM genera un archivo cuya única columna es un conjunto de números aleatorios, enteros o fraccionales, simulando su extracción de una distribución normal o uniforme.

4.2.1.2. Funcionamiento: Los pasos a seguir en la utilización de RANUM son los siguientes:

- (1) Teclee RANUM <CR> en la primera opción STATPAK
- (2) El programa pregunta por un número entero, en el intervalo (-32768, 32767) que se usa en el inicio de las rutinas generadoras de números aleatorios (se le denomina semilla). Digite el número y <CR >
- (3) Defina si requiere números enteros o reales (I,R)
- (4) Indique la distribución a usar para generar los números: normal o uniforme (N,E)
- (5) Descripción de los parámetros para la distribución seleccionada:
 - (A) Distribución normal: indique a RANUM la media y la desviación estándar deseadas
 - (B) Distribución uniforme: indique los límites inferior y superior para la distribución. Los límites se incluyen dentro de los posibles resultados
- (6) Teclee el número de valores que precisa
- (7) Defina el nombre del archivo de salida en el cual quedarán grabados los números aleatorios

4.2.1.3. Ejemplos:

Ejemplo (1): Selección de siete números enteros entre 00 y 99, para cierto juego de azar (!). En el cuadro 4.29

se dan las instrucciones, observe las siguientes anotaciones, marcadas con flechas:

- (1) Número inicial (semilla) 24478 [1]
- (2) Se quieren números enteros
- (3) Distribución uniforme (todos los números en el rango, tienen igual probabilidad de ser seleccionados)
- (4) Cada número debe pertenecer al intervalo [00,99]
- (5) Se requieren siete números aleatorios
- (6) El archivo de salida se denominará NUMALEAT.DAT
- (7) La salida se hace con "PROGRAM--SPEED 0", con lo cual se va al procedimiento SPEED con NUMALEAT.DAT como archivo de entrada
- (8) En SPEED se ordena imprimir el archivo de salida

Ejemplo (2): En una simulación se requiere una muestra de 50 observaciones pertenecientes a una distribución normal, con media 1.74 m y desviación estándar .09 m. El cuadro 4.31 muestra las instrucciones adecuadas en donde se puede apreciar:

- (1) El mismo número inicial (semilla) del anterior ejemplo
- (2) Se requieren números aleatorios reales y pertenecientes a una distribución normal...
- (3) ...Con media = 1.74 y desviación estándar = .09
- (4) Puesto que se da NUMALEAT.DAT como archivo de salida STATPAK avisa que este archivo ya existe (se creó en el anterior ejemplo) y pregunta si escribe encima de lo que ya contiene

[1] Nota: La misma semilla y los mismos parámetros distribucionales producirán siempre el mismo conjunto de números aleatorios

Cuadro 4.29

```

----> RANDOM NUMBERS <----

Random number seed (-32768 to 32767)? 24478
Integer or Real Numbers (I,R)? I
Normal or Even Distribution (N,E)? E
Lower Limit (inclusive)? 00
Upper Limit (inclusive)? 99
# of values in sequence? 7
Output File Name ('Q' to Quit): NUMALEAT.DAT
Working.....
Exit Command: <CR>=SP, R(erun), S(ystem), or PROGRAM -- SPEED 0

```

Cuadro 4.30

```

--> PROG:SPEED FILE:NUMALEAT.DAT DATE:02-27-1986 ID:PAGE 1

```

```


1: 55
2: 15
3: 92
4: 6
5: 71
6: 31
7: 78


```

En el cuadro 4.32 están los 50 números aleatorios generados por el programa. Posteriormente, en la sección 4.3.2. se utilizan estos números como ejemplo para el procedimiento FREHIST. Una mirada sobre esos resultados puede ayudar a tener una idea acerca de los conceptos de aleatorización y simulación.

Cuadro 4.31

----> RANDOM NUMBERS <----

Random number seed (-32768 to 32767)? 24478 

Integer or Real Numbers (I,R)? R 


Normal or Even Distribution (N,E)? N

Desired Mean? 1.74

Desired Standard Deviation? .09 

of values in sequence? 50

Output File Name ('Q' to Quit): NUMALEAT.DAT

NUMALEAT.DAT exists: re-use it (Y,N)? Y 

Working.....

Exit Command: <CR>=SP, R(erun), S(ystem), or PROGRAM --- SPED 0

Cuadro 4.32

--> PROG: SPEED FILE: NUMALEAT.DAT DATE: 02-27-1986 ID#PAGE 1

1:	1.645535
2:	1.671895
3:	1.673365
4:	1.566812
5:	1.64599
6:	1.790856
7:	1.721449
8:	1.779628
9:	1.757094
10:	1.749576
11:	1.729744
12:	1.804951
13:	1.715823
14:	1.565904
15:	1.849705
16:	1.68531
17:	1.70722
18:	1.704755
19:	1.70028
20:	1.680053
21:	1.627336
22:	1.780796
23:	1.655022
24:	1.772645
25:	1.74563
26:	1.756732
27:	1.773251
28:	1.609118
29:	1.72139
30:	1.693047
31:	1.66144
32:	1.647227
33:	1.622246
34:	1.813698
35:	1.751493
36:	1.661255
37:	1.75744
38:	1.674342
39:	1.635847
40:	1.783966
41:	1.778087
42:	1.684165
43:	1.893518
44:	1.807508
45:	1.808199
46:	1.745151
47:	1.768963
48:	1.787507
49:	1.54571
50:	1.872229

Cuadro 4.33 Resumen de los programas para manejo de distribuciones de probabilidad

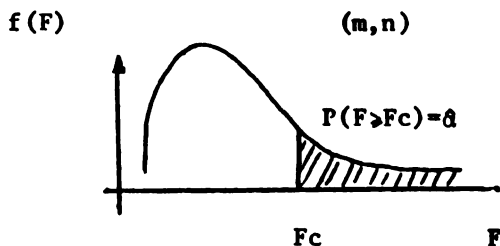
Programa	Clave	Distribución	Cálculos	Datos de Entrada
Condis 1	B	Normal Bivariada	Valor de la Función en un Punto (x,y)	$\rho_{xy}, \sigma_x, \sigma_y, \mu_x, \mu_y, x, y$
	F*	F de Snedecor	Probabilidad: $P(F > F_c)_{m,n} = \delta; E(F), V(F)$	Fc: FCalculada, m,n: Grados de Libertad
	G	Gamma	Probabilidad: $P(X < X_0)$	Xo, A:
	I	Gamma Incompleta	Probabilidad para pruebas de una o dos colas (δ)	t: t calculado, Grados de Libertad
	N	Normal (Inversa)		
	T*	T de Student		
.....				
Condis 2	C	Chi-cuadrada	Valor de la Función: $f(x_0)$ y Probabilidad: $P(X \leq x_0)$	$X_0 > 0$ y n: Grados de Libertad
	E	Exponencial	Valor de la Función: $f(x_0)$ y Probabilidad: $P(X \leq x_0)$	$X_0 > 0$ y E(x): Media de la Variable
	K	Khrgian-Mazin		
	L	Log-Normal		
	N*	Normal	Valor de la Función: $f(x_0), \frac{X_0 - \mu}{\sigma}$ Probabilidad: $P(X \leq x_0)$ y $z = \frac{X_0 - \mu}{\sigma}$	X_0, μ, σ
.....				
Discdis	B	Binomial	Probabilidad: $P(X = x_0)$ y $P(X \leq x_0), E(x), V(x)$	x_0, r, p
	H	Hipergeométrica	Probabilidad: $P(X = x_0)$ y $P(X \leq x_0), E(x), V(x)$	N° defectuosos, N° no defectuosos, tamaño de muestra Genera x.
	N	Binomial Negativa	Probabilidad: $P(x = x_0)$ y $P(X \leq x_0), E(x), V(x)$	Prob. de Exito, # de éxitos, # máximo de ensayos (x_0)
	P	Poisson	Probabilidad: $P(X = x_0)$ y $P(X \leq x_0)$	$x_0, E(x) = V(x) = \lambda = np$

4.2.2. CONDIS1, CONDIS2, DISCDIS

Distribuciones teóricas de probabilidad

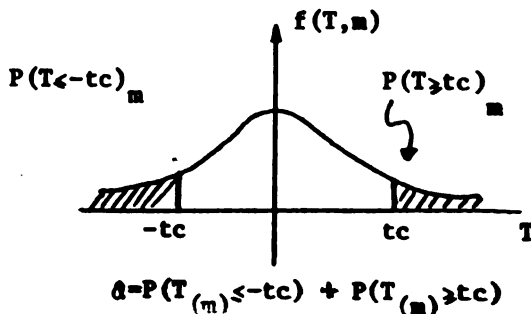
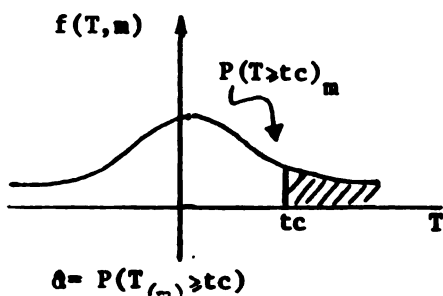
4.2.2.1. Descripción: Los tres programas realizan cálculos en distribuciones de probabilidad. Los dos primeros para variables aleatorias continuas, el tercero para discretas. El cuadro 4.33 presenta un resumen de las distribuciones que es posible manejar con estos programas, qué tipo de cálculo se puede hacer en cada una y cuáles son los valores de entrada que se deben indicar a la computadora. Las distribuciones marcadas con asterisco se describen en este manual; si el lector está interesado en alguna de las otras puede encontrar información en el manual de usuarios de STATPAK.

(1) Distribución F de SNEDECOR (F): calcula el áreas a la derecha de un valor dado (F calculado, en los casos prácticos). Esta área es $P(F > F_c)_{m,n}$, donde m y n son los grados de libertad de numerador y denominador respectivamente. Gráficamente:



En pruebas de hipótesis que utilizan esta distribución la $P(F > F_c)_{m,n}$ es la probabilidad de rechazar la hipótesis nula por error, esto es, la probabilidad estimada de cometer error tipo I (α) al tomar la decisión.

(2) Distribución T de Student (T): calcula el área debajo de la curva para pruebas de hipótesis de una o dos colas, dado un valor de T calculado y los grados de libertad (m) apropiados. Gráficamente:



Como en el caso anterior es la probabilidad, estimada, de cometer error tipo I al tomar la decisión en una prueba de hipótesis.

(3) Distribución normal (N): calcula el valor de la función de distribución, $f(x)$, el área bajo la curva a la izquierda, $P(X \leq x_i)$ y el valor estandarizado (z_i) para un conjunto de puntos x .

Sea la V.A $X \sim N(\mu, \sigma^2)$, la rutina obtendrá para cada valor x_0 , definido por el usuario, lo siguiente:

(i)

$$f(x_0) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x_0 - \mu}{\sigma}\right)^2}$$

(ii)

$$P(X \leq x_0) = \int_{-\infty}^{x_0} f(x) dx$$

(iii)

$$z_0 = \frac{x_0 - \mu}{\sigma}$$

El usuario tiene la opción de entrar cada valor x_0 o pedir al programa que genere valores en un intervalo y debe además indicar los valores para μ y σ .

Con los resultados de este programa se puede entonces graficar la distribución y/o hacer cálculo de probabilidades en ella.

4.2.2.2. Funcionamiento: las instrucciones requeridas para la utilización de cualquiera de los tres programas son las siguientes:

- (1) Teclee el nombre del programa de interés (CONDIS1, CONDIS2, DISCDIS) en la primera instrucción SP

- (2) Ya en el programa indique la distribución deseada, tecleando su letra correspondiente en el menú del programa que aparece en la pantalla
- (3) (a) Si utiliza distribución F de SNEDECOR indique a STATPAK el valor F calculado (x) y los grados de libertad del numerador (M1) y del denominador (M2)
- (b) Si trabaja con distribución T de Student defina a STATPAK el valor de t calculado (T) y los grados de libertad apropiados
- (c) En la distribución normal defina a STATPAK si los valores x_i están en un archivo o si se quiere que el programa los calcule automáticamente. En el primer caso indique también el nombre del archivo, en el segundo caso diga a STATPAK los valores inicial, final y de incremento para la generación de los números x_i . Luego entre los valores de la media y la desviación estándar de la distribución en cuestión

4.2.2.3. Ejemplos: Se plantean algunos problemas con base en ejemplos presentados anteriormente:

Ejemplo (1) Distribución F: en el primer conjunto de datos presentado en este manual se dieron los resultados de observaciones de porcentajes de azúcar y fibra en caña de azúcar. Se hizo un análisis de regresión $\% \text{ azúcar} = f(\% \text{ fibra})$, cuyos resultados aparecen en el cuadro 4.15, y que mostró un $F_c = 33.11874$ para el modelo, con 1 y 7 grados de libertad en numerador y denominador respectivamente. En el cuadro 4.34 se muestran las instrucciones y los resultados. Con flechas se marcan los siguientes puntos:

- (1) El menú del programa CONDISE
- (2) La selección F del menú y la respuesta de STATPAK
- (3) Entrada del valor F_c (x)
- (4) Definición de los grados de libertad del numerador (M1) y del denominador (M2)
- (5) El resultado: $= P(F \geq 33.11874)_{(1,7)}$
= .000694

Cuadro 4.34

```

----> CONTINUOUS DISTRIBUTION FUNCTIONS - 1 <----

B ivariate Normal Distribution
F Distribution
G amma Function
I ncomplete Gamma Function
N ormal Distribution (INVERSE)
t Distribution

X (exit this program)

Select desired function (B,F,G,I,N,T,X)? F

----> F Distribution <----

--> Enter values for requested variables
X? 33.11874
M1? 1
M2? 7

F(X,M1,M2) = 6.940963E-04
Variance = 7.84
Expected Mean Value = 1.4

Exit Command: <CR>=SP, R(erun), S(ystem), or PROGRAM --

```

(6) Varianza y media de la distribución F (1,7)

En este ejemplo el cálculo indica que si se decide aceptar que hay un efecto de dependencia ($H_a: \beta \neq 0$) del % fibra sobre el % caña, la probabilidad de que esta decisión esté errada es 0.069%. Así la decisión tiene un margen de seguridad de 99.93% lo cual conduce a tomarla.

Ejemplo 2. Distribución F: suponga que un diseño en bloques al azar dio los siguientes resultados:

F de	GL	CM	F_c
Bloques	4	300,15	
Tratamientos	5	657,24	2.16
Error	20	304,28	
TOTAL	29		

Es claro que las hipótesis sobre tratamientos son:

H_0 : Todos los efectos de los tratamientos son iguales

H_a : Al menos dos tratamientos difieren en su efecto poblacional

La decisión acerca de cual hipótesis es verdadera se toma con base en el valor F calculado. Tradicionalmente se compara F_c contra un F de tablas seleccionado con un $\alpha = .05$ o $\alpha = .01$, siendo α el nivel de significancia de la prueba y los grados de libertad apropiados (9 y 27 en este ejemplo). La facilidad que da el programa CONDISI, opción F es poder estimar la probabilidad de cometer error tipo I -rechazar H_0 cuando en la realidad H_0 es cierta-, α , para cada problema en particular.

En este caso los valores F de tablas son:

$$F(.05; 5, 20) = 2.71, \text{ mayor que } 2.16$$

$$F(.01; 5, 20) = 4.10, \text{ mayor que } 2.16$$

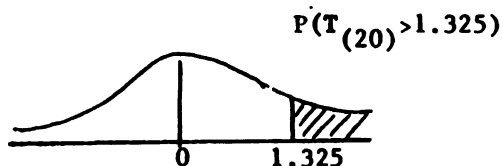
por lo cual la decisión del investigador siguiendo las normas tradicionales debe ser no rechazar la hipótesis nula (H_0), esto es, declarar que no hay evidencias de efectos diferenciados en los tratamientos.

Al utilizar CONDISI (cuadro 4.35) se obtiene que $\hat{\alpha} = P(F \geq 2.16)_{5,20} = .0997 = .10$, o 10%. Esto es: si se rechaza H_0 la probabilidad de que se haga por error es 10% ($\hat{\alpha}$). Ahora el investigador debe hacer un análisis profundo de su experimento, y quizá también de las consecuencias de la decisión que tome, antes de optar por rechazar o no la hipótesis nula. Lo nuevo ahora es que sabe cuál es la probabilidad de que se equivoque al tomar la decisión de que los tratamientos son poblacionalmente diferentes (10 en 100 en este caso) y la pregunta fundamental es: "solo se debe rechazar una hipótesis cuando la probabilidad de cometer error tipo I es .01 o .05?", reflexione un poco sobre ella, de su respuesta depende la utilidad de la opción F del programa CONDISI.

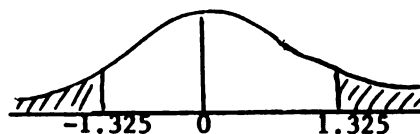
En el cuadro (4.35) se señalan: (1) los valores requeridos para el cálculo y (2) el resultado.

Ejemplo 3. Distribución T: puesto que la aplicación práctica del uso de T se estudiará en el programa MTESTS este ejemplo busca recordar un poco las propiedades de la distribución. El cuadro 4.36 muestra lo siguiente: en el caso marcado con la flecha 1 se da a STATPAK un valor calculado de 1.325 y un número de grados de libertad igual a 20 (señalados con (2)). La opción T de CONDISI calcula:

$$(3) P(T_{20} \leq -1.325) = P(T_{20} \geq +1.325) = .1001 \approx .10, \text{ gráficamente:}$$



$$(4) P(T_{20} \leq -1.325) \text{ o } (T_{20} \geq 1.325) = P(T_{20} \leq -1.325) + P(T_{20} \geq 1.325) = .10 + .10 = .20, \text{ gráficamente:}$$





$$(5) P(-1.325 \leq T_{20} \leq 1.325) = .7999 \approx .8 = 1 - .20, \text{ gráficamente}$$



El caso marcado con la flecha (6) es idéntico al anterior pero con $t_c = 2.086$

Cuadro 4.35

```
----> F Distribution <----  
--> Enter values for requested variables  
X? 2.16  
M1? 5   
M2? 20   
      F(X,M1,M2) = 9.977364E-02  
      Variance   = .7098766  
      Expected Mean Value = 1.111111  
Exit Command: <CR>=SP, R(erun), S(ystem), or PROGRAM --
```

Cuadro 4.36



----> t Distribution <---

--> Enter values for requested variables

Value of T? 1.325

Degrees of Freedom? 20



One-sided Tail Area = .1001



Two-sided Tail Area = .2001



Integral from -1.325 to 1.325 = .7999



Exit Command: <CR>=SP, R(erun), S(ystem), or PROGRAM --



----> t Distribution <---

--> Enter values for requested variables

Value of T? 2.086

Degrees of Freedom? 20

One-sided Tail Area = .0250

Two-sided Tail Area = .0500

Integral from -2.086 to 2.086 = .9500

Exit Command: <CR>=SP, R(erun), S(ystem), or PROGRAM --

Es interesante que el lector compare los valores obtenidos con las tablas de la distribución T de Student para una mejor comprensión del procedimiento.

Ejemplo 4. Distribución normal: un ejercicio sencillo. En el cuadro 4.37 se muestran las instrucciones para generar puntos y probabilidades en una distribución normal con media $\mu = 1.74$ y desviación estándar $\sigma = .09$. Observe las siguientes indicaciones señaladas con flechas en el cuadro:

- (1) El menú del programa CONDIS2.
- (2) Selección de la opción N y la respuesta de STATPAK
- (3) Se indica que el programa debe calcular los valores $x_i \dots$
- (4) ...comenzando con $x = 1.47$, terminando con $x = 2.01$ y aumentando .02 de valor en valor ($x_{i+1} = x_i + .02$)
- (5) Definición de media y desviación estándar de la distribución
- (6) Observe que los valores "Default" de la opción (5) son (0,1), esto es, que si no se dan valores para μ y σ , STATPAK trabaja con la distribución normal estándar (Z)

El cuadro 4.38 presenta los resultados:

- (1) Los valores generados de x_i
- (2) El valor de $f(x_i)$, ordenada de la curva,

$$f(x_i) = \frac{1}{(.09)\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x_i - 1.74}{.09}\right)^2}$$

- (3) La probabilidad acumulada para el punto $x_i: P(X \leq x_i)$, esto es, el área bajo la curva a la izquierda de x_i
- (4) Note que en la distribución normal definida: $X \sim N(1.74, .0081)$ la probabilidad de encontrar un individuo, en una selección al azar, con un valor entre 1.69 y 1.83 es:

$$P(1.69 \leq X \leq 1.83) = P(X \leq 1.83) - P(X \leq 1.69) = .8413 - .2892 = .5521$$

Cuadro 4.37



----> CONTINUOUS DISTRIBUTION FUNCTIONS - II <----

C hi-squared distribution
 E xponential distribution
 K hrgian-Mazin
 L ogarithmic-Normal distribution
 N ormal (Gaussian) distribution

X (eXit this program)

Select desired function (C,E,K,L,N,X)? N

----> Normal Distribution <----

File or Automatic X-values (F,A)? A

Initial Value of X ? 1.47

Final Value of X ? 2.01

Step Value of X ? .02

Mean of Distribution (<CR>=0): 1.74

S.D. of Distribution (<CR>=1): .09

Output : S(creen), P(rinter), F(ile), or Q(uit)?

Cuadro 4.38

--> PROG:CONDIS2 DATE:02-27-1986 ID:Normal

X	DENSITY f(x)	CUMULATIVE F(x)	NORMALIZED x
1.47	4.431856E-03	1.349968E-03	-3
1.49	8.421542E-03	2.736659E-03	-2.777778
1.51	.0152318	5.300944E-03	-2.555556
1.53	2.622188E-02	9.815287E-03	-2.333334
1.55	4.296652E-02	1.738128E-02	-2.111112
1.57	6.701156E-02	.0294532	-1.88889
1.59	9.947701E-02	4.779014E-02	-1.666668
1.61	.140556	7.430672E-02	-1.444446
1.63	.1890293	.1108113	-1.222224
1.65	.2419704	.1586544	-1.000002
1.67	.2948146	.2183487	-.777796
1.69	.341892	.2892557	-.555577
1.71	.3773831	.3694391	-.333357
1.73	.3964873	.4557611	-.111137
1.75	.3964876	.5442369	.1111084
1.77	.3773839	.630559	.3333304
1.79	.3418931	.7107426	.5555524
1.81	.2948158	.7816497	.7777743
1.83	.2419717	.8413442	.9999964
1.85	.1890305	.8891876	1.222218
1.87	.140557	.9256926	1.44444
1.89	9.947789E-02	.9522094	1.666662
1.91	6.701224E-02	.9705464	1.888884
1.93	4.296699E-02	.9826185	2.111107
1.95	.0262222	.9901846	2.333329
1.97	.015232	.994699	2.55555
1.99	8.421665E-03	.9972633	2.777773
2.01	4.431928E-03	.9986499	2.999994

Así que el 55.2% de los individuos de esta población toman valores para la variable X en el intervalo (1.69, .83).

Suponga que X mide la estatura de personas del sexo masculino en Uruguay. Calcule algunas probabilidades de interés para recordar (!)

- (5) Finalmente se da el valor estandarizado (Z) para cada punto.

$$z_1 = \frac{x_1 - 1.74}{.09}$$

Con el fin de ver un poco mejor los resultados el cuadro 4.39 presenta la gráfica $X = x_1$, $Y = f(x_1)$ obtenida con el programa QPOLY para los datos del problema. Trate de hacerla, llamando a QPLOT y dejándose guiar por las instrucciones del programa... y el sentido común.

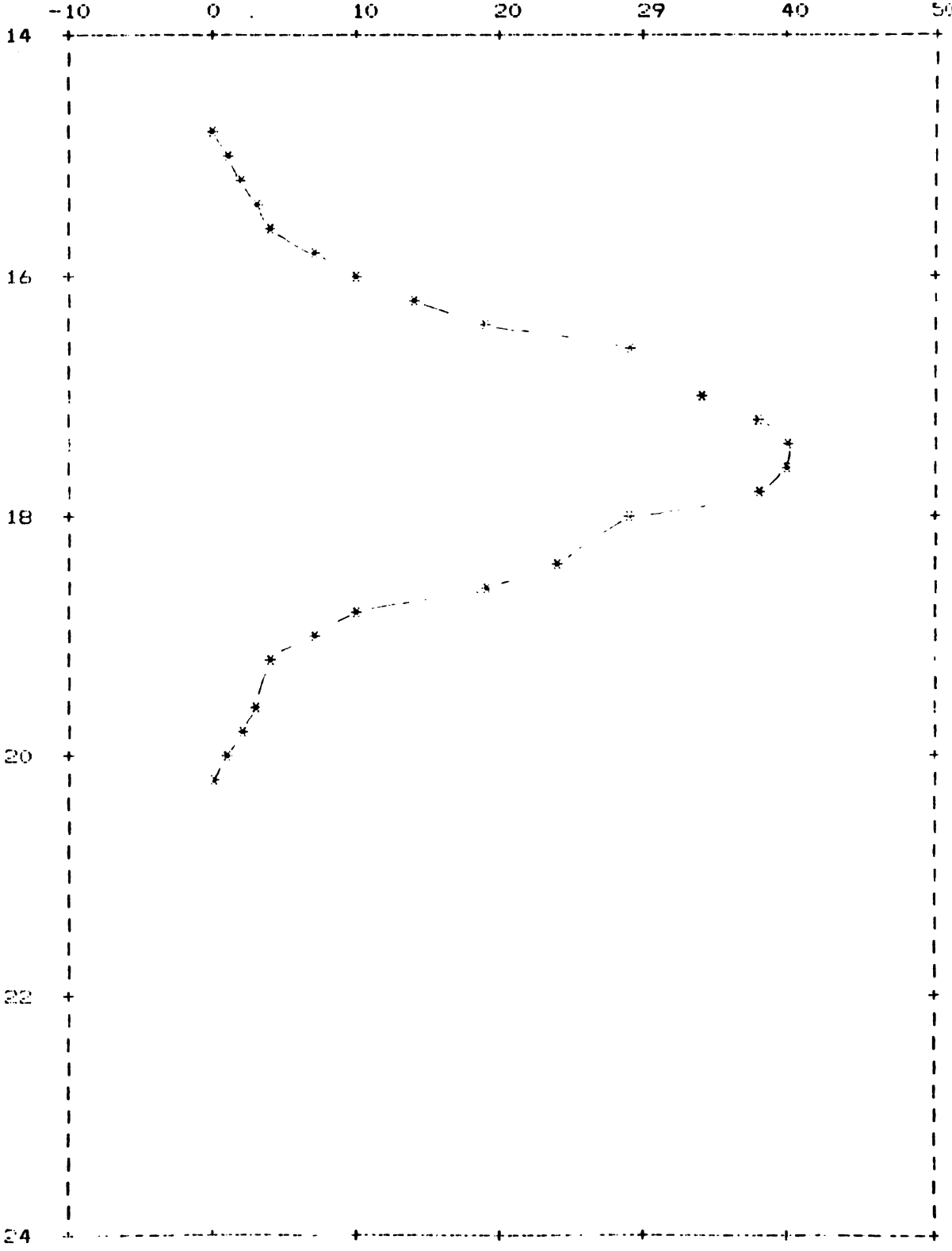
X * .1

Y * .01

COLLISION MARK: ?

FILE: NOR.DAT

Columns: X = 1 (*) = 2



4.3 Programas para la descripción estadística de muestras y poblaciones

4.3.1 DSTATS

Cálculo de Estadísticas Descriptivas

4.3.1.1. Descripción: DSTATS calcula un conjunto de estadísticas descriptivas para cada una de un grupo de variables medidas en una muestra o población. Estas estadísticas son:

Mínimo
 Máximo
 Media aritmética
 Desviación estándar
 Varianza
 Error estándar (desviación estándar de la media)
 Coeficiente de variación
 Media armónica
 Media cuadrática (RMS)
 Media Geométrica
 Tercer momento alrededor de la media
 Cuarto momento alrededor de la media
 Coeficiente de falta de simetría (Skewness)
 Coeficiente de achatamiento (Kurtosis)

4.3.1.2. Funcionamiento: las instrucciones necesarias para ejecutar este programa son las siguientes:

- (1) Teclee DSTATS <CR> en la primera instrucción SP
- (2) Indique al programa si los datos corresponden a una población (P) o muestra (S: sample)
- (3) Defina el número de variables (columnas-campos) que requiere analizar
- (4) Realice la asignación de campos
- (5) Indique el tipo de salida que desea

4.3.1.3. Ejemplo: en la sección 4.1.3. del manual se introdujo un ejemplo del libro de STEEL and TORRIE, correspondiente a experimentación en tabaco. Luego de hacer el análisis de regresión múltiple se generó el archivo TABACO.RSD que se presenta en el cuadro 4.40, el cual contiene los valores observados (y) en la columna 1, los estimados con la ecuación de regresión (\hat{y}) en la columna 2 y los residuales ($y - \hat{y}$) en la tercera columna, para las 30 observaciones. Se trabaja sobre estas variables en la ejemplificación de DSTATS.

Cuadro 4.40

--> PROG: SPEED FILE: TABACO.RSD DATE: 03-01-1986 ID: PAGE 1

1:	.34	.805031	-.455031
2:	.11	3.899404E-02	7.100596E-02
3:	.58	.7929974	-.4129974
4:	.68	.6030695	7.693053E-02
5:	.18	.2179599	-3.795987E-02
6:	0	.2520875	-.2520875
7:	.08	3.787763E-02	4.212237E-02
8:	.11	.3298776	-.2198776
9:	1.53	1.206022	.3239776
10:	.77	1.146177	-.3761772
11:	1.17	1.097135	7.286525E-02
12:	1.01	1.051564	-4.156397E-02
13:	.89	1.269597	-.3795972
14:	1.4	1.195504	.2044963
15:	1.05	.9840683	6.593115E-02
16:	1.15	1.148933	1.067042E-03
17:	1.49	1.377305	.1126947
18:	.51	.4293526	8.064741E-02
19:	.18	.1659392	1.406081E-02
20:	.34	.4364223	-9.642226E-02
21:	.36	.1752021	.1847979
22:	.89	.8823537	7.646323E-03
23:	.91	1.036143	-.1261432
24:	.92	.7116506	.2083495
25:	1.35	1.071237	.2787623
26:	1.33	1.125642	.204358
27:	.23	.2801342	-.0501342
28:	1.26	.502196	.7578041
29:	.73	.8048837	-.0748837
30:	.23	.4046495	-.1746495


El cuadro 4.41 muestra las instrucciones utilizadas para el cálculo de las estadísticas simples, marcadas con flechas se dan las siguientes observaciones:

- (1) Título del programa
- (2) Los datos son una muestra (S)
- (3) Se desea procesar dos columnas ...
- (4) ... Del archivo TABACO.RSD (<CR>)
- (5) Las dos columnas a procesar son la segunda (y) y la tercera (y - \hat{y})
- (6) Salida a la impresora






Los resultados se presentan en los cuadros 4.42 y 4.43 y corresponden a la salida, identificada, por la impresora.

En caso de que se indique mostrar los resultados en la pantalla (S), cuadro 4.44, la máquina presenta los resultados para la primera variable (columna 2) y al final la palabra Holding .., indicando que hay más información. Tecleando <CR> mostrará el análisis para la segunda variable. Observe los puntos mencionados señalados con flechas en el cuadro.

Cuadro 4.41



 ---- DESCRIPTIVE STATISTICS ----

Sample or Population Statistics (S,P)? S 
 # of Fields to process (NCR=1)? 2 
 Input File (Q=Quit, <CR>=TABACO.RSD):  TABACO.RSD
 2 Data Field(s)
 3 File Columns in TABACO.RSD
 Field Assignment(s) (NCR=1:1): 2 3 
 Output : S(screen), P(rinter), F(ile), or Q(uit)? P 

Cuadro 4.42

--> PROG: DSTATS FILE: TABACO.RSD DATE: 03-01-1986 ID: ESTADISTICAS SIMPLES PARA VALORES ESPERADOS Y RESIDUALES
PAGE 1

* = SAMPLE Statistics

FILE COLUMN = 2

Data Elements (N): 30 Sum: 21.58091 Min: 3.787763E-02 Max: 1.377305

- * Arithmetic Mean: .7193356
- * Standard Deviation: .411734
- * Variance: .1701702
- * Standard error: 7.575624E-02
- * Coefficient of Variation (%): 57.63311

- Harmonic mean: .2815609
- Quadratic mean (RMS): .8259536
- Geometric mean: .5262599
- * Third Moment about the mean: -1.155161E-02
- * Fourth moment about the mean: 4.665344E-02
- * Coefficient of skewness: .1614184
- * Coefficient of kurtosis: 1.577865

--> PROG: DSTATS FILE: TABACO.RSD DATE: 03-01-1986 10: ESTADISTICAS SIMPLES PARA VALORES ESPERADOS
PAGE 2

Y RESIDUALES

* = SAMPLE Statistics

FILE COLUMN = 3

Data Elements (N): 30 Sum: -6.765127E-06 Min: -.465031 Max: .7578041

Arithmetic Mean: -2.255043E-07
 * Standard Deviation: .2499138
 * Variance: 6.245688E-02
 * Standard error: .0456278
 * Coefficient of Variation (%): -1.198244E+08


Harmonic mean: 2.627109E-02
 Quadratic mean (RMS): .2457132
 Geometric mean: .1030005
 * Third Moment about the mean: 7.321405E-03
 * Fourth moment about the mean: 1.648085E-02
 * Coefficient of skewness: .469055
 * Coefficient of kurtosis: 4.225693

Cuadro 4.44

Sample or Population Statistics (S,P)? 5
 # of Fields to process (CORP=1)? 2
 Input File (Default, <CRP=FABACO.RSD):
 2 Data Field(s)
 3 File Columns in FABACO.RSD
 Field Assignment(s) (CUR=1): 2 3

Output : 3(screen), P(printer), F(file), or Q(quit) 5

* - SAMPLE Statistics

FILE COLUMN = 2 

Data Elements (N): 00 Sum: 21.50001 Min: 3.787765E+02 Max: 1.377305

- * Arithmetic Mean: .7195336
- * Standard Deviation: .414734
- * Variance: .1721702
- * Standard error: 7.575624E-02
- * Coefficient of Variation (%): 57.5831.

- Harmonic mean: .2615609
- Quadratic mean (RMS): .9269526
- Geometric mean: .5062569
- * Third Moment about the mean: -1.195161E-02
- * Fourth moment about the mean: 4.663344E-02
- * Coefficient of skewness: -.1616194
- * Coefficient of kurtosis: 1.573860

Holding... 

4.3.2. FREHIST

Histogramas y Tablas de Frecuencias

4.3.2.1. Descripción: este programa realiza análisis de frecuencias sobre conjuntos de datos numéricos, cuyos resultados son presentados en forma numérica (tabla) o gráfica (histograma). Funciona interactivamente, dando al usuario la opción de definir los parámetros del análisis (mínimo, máximo y longitud del intervalo), o permitir que el programa los defina. Da, también, algunas alternativas para la presentación de resultados, tales como adicionar las frecuencias relativas (en porcentaje) y/o la ojiva de frecuencias acumuladas, al histograma.

4.3.2.2. Funcionamiento: el orden secuencial de trabajo con FREHIST es el siguiente:

- (1) Entre FREHIST <CR> en la primera instrucción STATPAK
- (2) El programa preguntará por la definición del archivo de entrada y la asignación de campos
- (3) Indique si la computadora debe buscar mínimo y máximo de los datos, para usarlos como límite inferior y superior, o si estos límites van a ser proporcionados al programa por el usuario
- (4) Defina los límites inferior y superior si escogió esa opción
- (5) Instrucción de definición del histograma

El programa presenta un conjunto de instrucciones previas a la construcción del histograma con el objeto de definir los parámetros necesarios: límites inferior y superior, número de intervalos y ancho de intervalo. Hay tres posibles acciones para el usuario:

- (1) Especificar alguno o algunos de los valores
- (ii) Escoger como valores los ofrecidos por el programa
- (iii) Pedir al programa que calcule y use los valores calculados

En el primer caso se da el valor y <CR> en el segundo solo <CR> y en el tercero <?><CR.> Los valores ofrecidos por el programa para los límites inferior y superior son (0,100) si no se pidió en el paso (3) que buscara mínimo y máximo de los datos o (mín, máx) si se le pidió el cálculo en el paso 3. Para el número de intervalos y longitud de intervalo el valor ofrecido es 10.

- (6) Terminado el paso anterior la computadora pregunta si el histograma requerido es el descrito en pantalla, dando la opción de hacer correcciones
- (7) Indique si desea salida numérica (tabla de frecuencia) o gráfica (histograma)
- (8) En caso de pedir salida gráfica indique al programa
 - (i) el ancho de hoja que está utilizando (si no lo da STATPAK asume 65, esto es, hoja tamaño carta)
 - (ii) que adicione, o no, a cada barra en el histograma la frecuencia acumulada y escriba su valor
 - (iii) que escriba los porcentajes (frecuencia relativa) de cada clase
- (9) Defina a la máquina el tipo de salida para los resultados

4.3.2.3. Ejemplo: se utilizan los datos presentados en el cuadro 4.40 para hacer un análisis de frecuencias a los valores residuales de la regresión. Es importante recordar que el método de mínimos cuadrados hace ciertas suposiciones sobre los términos aleatorios del error experimental y que una observación del histograma de frecuencias de los residuales da alguna idea acerca de esas suposiciones.

En los cuadros 4.45 y 4.46 se señalan, con flechas numeradas, algunos puntos interesantes sobre el ejemplo y que se describen en adelante:

- (1) El análisis se hace a la tercera columna (residuales) del archivo TABACO.RSD

Cuadro 4.45

----> FREQUENCY HISTOGRAM <----

Input File (Q=Quit, <CR>=TABACO.RSD):

1 Data Field(s)

3 File Columns in TABACO.RSD

Field Assignment(s) (<CR>=1:1): 3

Determine Minimum & Maximum (Y/N)? Y

Finding Limits.....

30 Data elements

Minimum Value = -.465031

Maximum Value = .7578041

----> Histogram Definition <----

Lower Limit (<?>= Calc, <CR>=-.465031):

Upper Limit (<?>= Calc, <CR>= .8578041): .7578041

of Intervals (<?>= Calc, <CR>= 10): 5

Histogram: -.465031 to .7578041 in 5 intervals .244567 wide -

OK (Y/N)

Cuadro 4.46

PRG: FREHIST FILE: TABACU.RSD DATE: 03-01-1986 ID: TABLA DE FRECUENCIAS PARA LOS RESIDUALES

PAGE 1

FROM	TO BELOW	FREQ	%	CUMUL	%
-.465031	-.220464	5	17.2	5	17.2
-.220464	2.410302E-02	11	37.9	16	55.1
2.410302E-02	.26867	11	37.9	27	93.1
.26867	.513237	2	6.8	29	100
.513237	.7578041	0	0	29	100

MEDIAN

PRG: FREHIST FILE: TABACU.RSD DATE: 03-01-1986 ID: HISTOGRAMA DE FRECUENCIAS PARA LOS RESIDUALES

PAGE 1

LOW LIMIT (DX = .244567 SCALE = 1:1)

```

-.465031 |*****
-.220464 |*****M*****
2.410302E-02 |*****
.26867 |**
.513237 |
    
```

HISTOGRAM STATISTICS

Mean -1.384705E-02
 Variance .0423983
 S.D. .2058342
 S.D. of Mean 3.75914E-02
 Skewness -2.54911E-02

Low Outliers = 0 High Outliers =

- (2) Se opta porque el programa encuentre mínimo y máximo de los datos, el programa busca y presenta esos valores
- (3) Definición del histograma: note que no se define el ancho de los intervalos puesto que al fijar límites inferior y superior y el número de intervalos el ancho queda definido implícitamente. Otro punto interesante a observar es que al fijar como límite superior del histograma el valor .7578041 (el máximo) se eliminó este valor de la tabla (ver punto 5) quedando solo 29 valores en total y 0 observaciones en la última clase
- (4) El programa pregunta si se aceptan las definiciones hechas para el análisis
- (5) Los límites de intervalo son de la forma [,) esto es [mayor o igual, menor estrictamente)
- (6) La clase que contiene la mediana se identifica, en este caso es la segunda con límites [-.220464, .02410302)
- (7) Salida gráfica
- (8) Escala 1:1, indicando que cada signo = o * representa un dato
- (9) Estadísticas descriptivas, calculadas para la tabla de frecuencias (fórmulas para datos agrupados)
- (10) El programa indica que un valor quedó fuera, por arriba, del histograma. En este ejemplo ese valor es el máximo (ver punto 3)

Finalmente, para completar el manejo del archivo de residuales, se presenta, en el cuadro 4.47, la gráfica de valores esperados vs residuales utilizable para el llamado "análisis de residuales" usado en el estudio de los supuestos del método de mínimos cuadrados.

4.3.3. FREQUICK

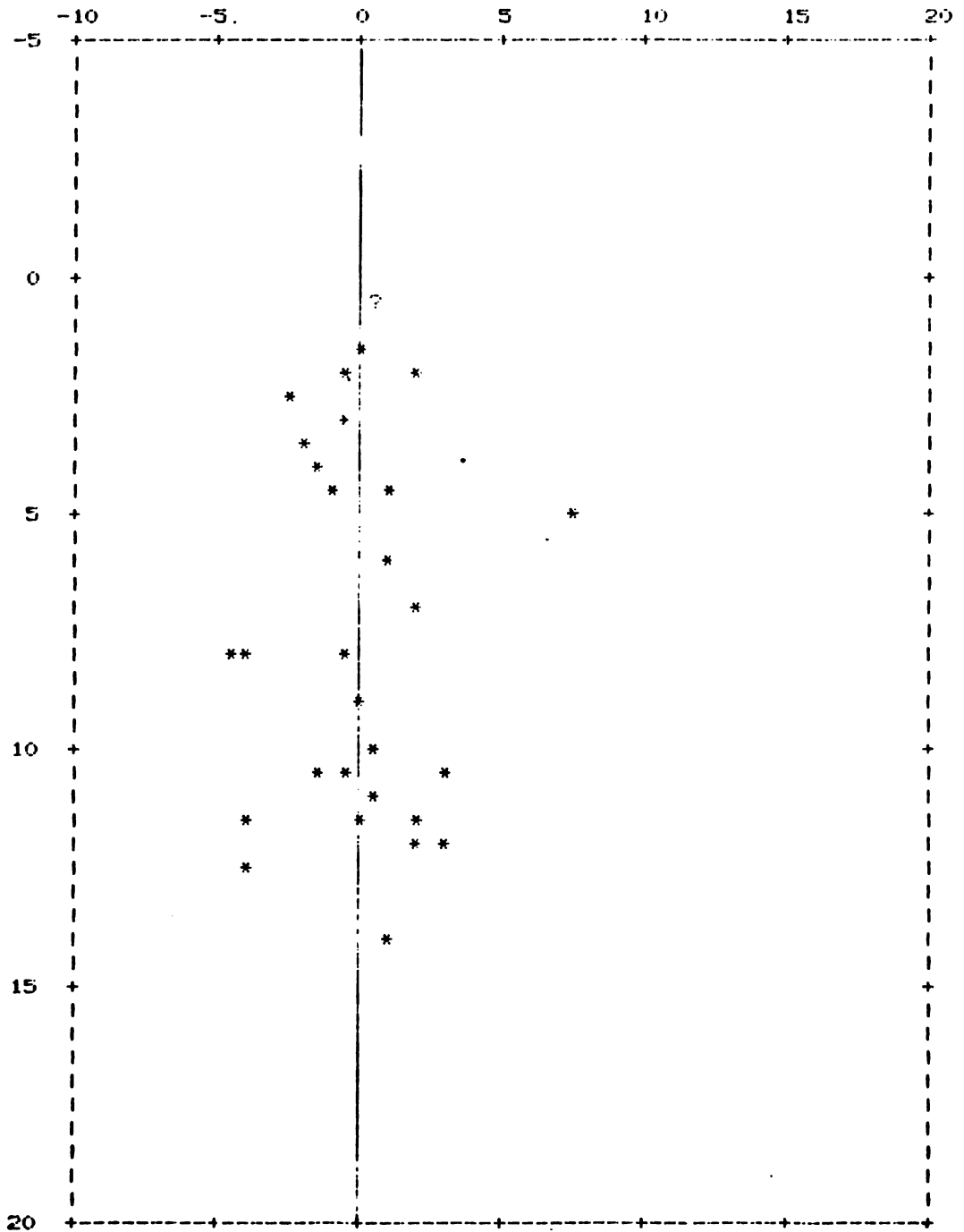
Análisis rápido de frecuencias para datos discretos

4.3.3.1. Descripción: FREQUICK realiza un conteo de los diferentes valores que presenta una variable de interés, estos

ID: GRAFICA DE VALORES ESPERADOS VS. RESIDUALES Date: 03-01-1986

X * .1 Y * .1 COLLISION MARK: ?

FILE: TABACO.SRT Columns: X = 1 (*) = 2



valores pueden ser numéricos o alfabéticos y cada uno de ellos forma un intervalo con el fin de presentar los resultados del conteo en la forma de una tabla o un histograma de frecuencias. Las variables numéricas continuas se toman como discretas para el análisis.

4.3.3.2. Funcionamiento: los pasos de funcionamiento son los siguientes:

- (1) Entre la palabra FREQUICK <CR> en la primera instrucción STATPAK
- (2) Defina el archivo sobre el cual quiere trabajar y la asignación de campos correspondiente
- (3) Indique el tipo de salida deseado (tabla o histograma)
- (4) En caso de pedir gráfica (histograma) se debe definir el ancho de hoja e indicar al programa si se quiere que adicione las frecuencias acumuladas y/o los porcentajes correspondientes a la frecuencia relativa de cada clase

4.3.3.3. Ejemplo: suponga que en una muestra se tienen 14 observaciones de estatura (m) clasificadas según el sexo de la persona (M: mujer; H: hombre), como se muestra en el cuadro 4.48, y se quiere contar y calcular la proporción de hombres y mujeres. En los cuadros 4.49.a y 4.49.b y 4.49.c se muestran las instrucciones dadas al programa y los resultados obtenidos con ellas. Marcadas con flechas algunas anotaciones de interés:

- (1) El análisis se hace a la primera columna del archivo FREC.DAT
- (2) Se pide una salida numérica por la impresora
- (3) Se identifica la salida con un título
- (4) La máquina imprime los resultados bajo un encabezado con el nombre del programa, el archivo de datos, la fecha del análisis y la identificación. Los resultados indican que hay 6 hombres (42.8%) y 8 mujeres (57.1%)
- (5) En el segundo caso se le pide al programa una salida gráfica (histograma) con longitud de línea de 65, la frecuencia acumulada adicionada al gráfico y sin porcentajes

- (6) El gráfico se hizo con escala 1:1
- (7) Aparecen las seis observaciones de la primera clase en la segunda (frecuencia acumulada)
- (8) En el tercer caso se pide una salida gráfica con ancho de hoja 80, frecuencias acumuladas sobreimpuestas y con la inclusión de los porcentajes
- (9) Observe la inclusión de los porcentajes

Cuadro 4.48

--> PROG: SPEED FILE: FREC.DAT DATE: 03-31-1986 ID: PAGE 1

1:	M	1.65
2:	H	1.68
3:	H	1.70
4:	M	1.70
5:	M	1.64
6:	M	1.67
7:	H	1.71
8:	H	1.80
9:	M	1.75
10:	M	1.66
11:	H	1.85
12:	H	1.78
13:	M	1.65
14:	M	1.64

Cuadro 4.49 a

Input File (Q=Quit, <CR>=FREC.DAT):
 1 Data Field(s)
 2 File Columns in FREC.DAT
 Field Assignment(s) (<CR>=1:1): 1

Counting.....

2 distinct cells.

Sorting....

Output : S(screen), P(printer), F(file), or Q(quit)? P
 Numeric or Graphic output (N/G)? N

--> MAKE PRINTER READY
 Report ID: EJEMPLO SOBRE EL ANALISIS RAPIDO DE FRECUENCIAS

--> PROG:FREQUICK FILE:FREC.DAT DATE:03-31-1986 ID:EJEMPLO SOBRE EL ANALISIS RAPIDO DE FRECUENCIAS
 PAGE 1

#	CELL VALUE	FREQUENCY	%	CUMULATIVE	%
1	H	6	42.8	6	42.8
2	M	8	57.1	14	100.0

Cuadro 4.49 b





---> Quick Frequency Count. <-T-

Input File (Q=Quit, <CR>=FREC.DAT):
1 Data Field(s)
2 File Columns in FREC.DAT
Field Assignment(s) (<CR>=1:1): 1

Counting.....

2 distinct cells.

Sorting.....

Output : S(screen), P(rinter), F(ile), or Q(uit)? P
Numeric or Graphic output (N/G)? G 
Maximum Histogram Line Length (<CR>=65)? 
Superimpose Cumulative Frequency (Y/N)? Y 
Include Percentages (Y/N)? N 

--> MAKE PRINTER READY
Report ID: SALIDA GRAFICA DEL PROGRAMA FREQUICK

---> PROG:FREQUICK FILE:FREC.DAT DATE:03-31-1986 ID:SALIDA GRAFICA DEL PROGRAMA FREQUICK
PAGE 1

CELL VALUE (SCALE = 1:1)

H |*****
M |*****+-----



Cuadro 4.49 c

----> Quick Frequency Count <----

Input File (Q=Quit, <CR>=FREC.DAT):

1 Data Field(s)

2 File Columns in FREC.DAT

Field Assignment(s) (<CR>=1:1): 1

Counting.....

2 distinct cells.

Sorting.....

Output : S(screen), P(printer), F(file), or Q(quit)? P

Numeric or Graphic output (N/G)? G

Maximum Histogram Line Length (<CR>=65)? 80

Superimpose Cumulative Frequency (Y/N)? Y

Include Percentages (Y/N)? Y

--> MAKE PRINTER READY

Report ID: FREQUIK, CAMBIANDO ANCHO Y CON PORCENTAJES

--> PROG:FREQUIK FILE:FREC.DAT DATE:03-1-1986 ID:FREQUIK, CAMBIANDO ANCHO Y CON PORCENTAJES
PAGE 1

CELL VALUE < SCALE = 1:1 >

H |***** <42.8%>
M |*****+----- <100.0%>

4.4 Pruebas de hipótesis que utilizan la distribución "T de Student"

4.4.1. MTESTS

Pruebas de hipótesis T de Student

4.4.1.1. Descripción: MTESTS es un conjunto de programas que tienen en común la utilización de la distribución de T. Los programas son: (ver cuadro 4.50)

(1) Prueba de T para el coeficiente de correlación

Obtiene el valor t_c (t - calculado) para la hipótesis $H_0: \rho = 0$, el nivel de significancia α para las hipótesis alternas $H_a: \rho > 0$, $H_a: \rho < 0$ (prueba de una cola) y $H_a: \rho \neq 0$ (prueba de dos colas) y un intervalo de confianza de 95% para el coeficiente de correlación poblacional. Adicionalmente obtiene el valor t_c (t - calculado) para la hipótesis nula $H_0: \rho = k$ (k es un valor propuesto por el usuario) y la significancia para las hipótesis alternas $H_a: \rho < k$, $H_a: \rho > k$ y $H_a: \rho \neq k$

(2) Comparación de las medias de dos poblaciones, muestras independientes:

calcula el valor t_c para la prueba de la hipótesis nula $H_0: \mu_1 - \mu_2 = k$ donde k es una constante (usualmente cero) y μ_1 , μ_2 las medias poblacionales, desconocidas. El usuario debe indicar si las varianzas de las poblaciones se asumen iguales o diferentes. Además del cálculo de t_c el programa calcula los niveles de significancia para las hipótesis alternas $H_a: \mu_1 - \mu_2 > k$, $H_a: \mu_1 - \mu_2 < k$ y $H_a: \mu_1 - \mu_2 \neq k$. El

Cuadro 4.50

```
---> MEANS TESTS <---  
  
C orrelation Coefficient t-statistic  
O rdinary t-statistic, Unpaired Samples  
P earson Product Moment Correlation  
S ingle Sample t-statistic  
t -statistic, Paired Samples  
  
X (eXit this program)  
  
Select desired function (C,U,P,S,T,X)? C
```

programa asume que las muestras son independientes (o "no pareadas")

- (3) Hipótesis sobre la media de una población: obtiene el valor t_c para la hipótesis $H_0: \mu = k$, donde k es una constante elegida por el usuario, los valores de significancia para hipótesis alternas de una ($H_a: \mu < k$, $H_a: \mu > k$) o dos colas ($H_a: \mu \neq k$). El usuario indica si se asume varianza poblacional conocida o desconocida
- (4) Comparación de las medias de dos poblaciones, muestras pareadas: calcula el valor t_c para la hipótesis nula $H_0: \mu_1 = \mu_2$ donde μ_1 y μ_2 son las medias poblacionales desconocidas, y la significancia para las hipótesis alternas $H_a: \mu_1 < \mu_2$, $H_a: \mu_1 > \mu_2$ y $H_a: \mu_1 \neq \mu_2$. Asume que las muestras son pareadas, esto es que hay una fuente de variación adicional que se debe eliminar (corresponde al diseño de bloques al azar con dos tratamientos)

4.4.1.2. Funcionamiento: al teclear en la primera instrucción STATPAK la palabra MTESTS <CR> aparece en la pantalla la lista de programas disponibles (cuadro 4.50) los cuales pueden ser llamados con una letra clave. La opción P (coeficiente de correlación de Pearson) se verá adelante cuando se toque ese tema específico. Los pasos a seguir posteriormente dependen del programa elegido y se tratarán uno por uno con base en los ejemplos.

4.4.1.3. Ejemplos

4.4.1.3.1. Prueba de t para el coeficiente de correlación (cuadro 4.51)

- (1) Al teclear en la pantalla inicial C <CR> aparece el

título y la primera pregunta

- (2) Se define el número de observaciones (parejas x , y), en este caso 30
- (3) Se indica el coeficiente de correlación estimado, en este caso $-.4163$
- (4) Si se quiere probar $H_0: \rho = k$ se define el valor de k en el intervalo $[-1,1]$, en este caso no se da un valor por lo cual el programa asume $k = 0$
- (5) Valor de t_c para $H_0: \rho = 0$

Resultados

- (6) Niveles de significancia para hipótesis alternas de una y dos colas. En este caso los resultados son:
- (i) la probabilidad de equivocarse $\hat{\alpha}$ al rechazar $H_0: \rho = 0$ y aceptar $H_a: \rho < 0$ o $H_a: \rho > 0$ es $= .0109$ o 1% lo cual conduce a pensar que se puede rechazar H_0 con una alta seguridad (99%)
- (ii) la probabilidad de equivocarse, $\hat{\alpha}$, al rechazar $H_0: \rho = 0$ y aceptar $H_a: \rho \neq 0$ es $= .0219$ (2%)

Es claro entonces que el programa estima la probabilidad de cometer error tipo I al hacer la prueba de hipótesis

Cuadro 4.51

```

----> Correlation Coefficient Test Statistic <----
# of data items (N)? 50
Estimate of Correlation Coefficient (r)? -.4163
Hypothesized value for r(0) (COR=0)?
t-statistic for r=0: -2.422774
One-sided Significance Level = .0106
Two-sided Significance Level = .0213
approximate 95% C.I. [-0.6753 to 0.06371
z-statistic for hypothesis: -2.002776
One-sided Significance Level = .0106
Two-sided Significance Level = .0213
Exit Command: CTRL=CF, R(erun), S(ystem), or PROGRAM -- R

```


- (7) Intervalo de confianza del 95% para el coeficiente de correlación poblacional
- (8) Valor de t_c para $H_0: \rho = k$, en este caso $k = 0$
- (9) Niveles de significancia para hipótesis alternas de una y dos colas. Note que, salvo errores de redondeo, son los mismos valores obtenidos en (6), esto se debe a que se definió $k = 0$
- (10) Salida del programa. En este caso se pide que vuelva a entrar en MTESTS pues se da la opción $R < CR >$

4.4.1.3.2. Comparación de dos poblaciones muestras independientes

El cuadro 4.52 muestra un conjunto de datos presentado en el libro de Steel and Torrie, página 76, y que corresponden al estudio del efecto de la vitamina A sobre la ganancia de peso de vaquillonas holando. La primera columna contiene las ganancias del grupo de animales testigo (sin aplicación de vitamina A), la segunda contiene ganancias de peso de animales tratados con vitamina A. Las muestras se asumen independientes puesto que se tomaron 28 animales homogéneos y se asignaron a los tratamientos aleatoriamente.

El cuadro 4.53 presenta las instrucciones dadas a la computadora y los resultados obtenidos; con flechas se señalan los siguientes puntos:


- (1) Título del programa, aparece en la pantalla al teclear 0 <CR> en la lista inicial de opciones
- (2) Se indica al programa asumir varianzas "poblacionales" iguales, se supone que se ha hecho una prueba previa
- (3) El programa da la opción de probar $H_0: \mu_1 - \mu_2 = k$, en este caso se escoge $k = 0$, esto es se prueba la

Cuadro 4.52

--- PROG: SPEED FILE: TINDEX.DAT DATE: 03-01-1986 IL: PACH 1


1: 175 142
 2: 132 311
 3: 218 337
 4: 151 262
 5: 200 302
 6: 219 195
 7: 234 253
 8: 149 199
 9: 187 236
 10: 123 216
 11: 248 211
 12: 206 176
 13: 179 249
 14: 206 214

Cuadro 4.53

--- Unpaired Sample t-statistic --- 

Do X & Y have the same Variance (Y/N)? Y 


Test value for Null Hypothesis (CUR=0) 

--- X-ARRAY --- 


Input File (Q=Quit, <CR>=TINDEX.DAT):


1 Data Field(s)

2 File Columns in TINDEX.DAT

Field Assignment(s) (<CR>=1:1): 1 

Reading data.....


14 X-elements 

--- Y-ARRAY --- 

Input File (Q=Quit, <CR>=TINDEX.DAT):

1 Data Field(s)

2 File Columns in TINDEX.DAT

Field Assignment(s) (<CR>=1:1): 2 

Cuadro 4.53 (continuación)

2 File Columns in TIMEP.DAT
 Field Assignment(s) (CUR)=1:1): 2
 Reading Data.....
 14 Y-elements ← 7

t-statistic = -2.724155 ← 8

One-sided Significance Level = .0057 ← 9
 Two-sided Significance Level = .0114

X-ARRAY ---	Mean =	157.6429	SD =	38.07827	← 10
Y-ARRAY ---	Mean =	275.9280	SD =	54.29623	

Difference of the Means	=	48.28572	
Std. Error of the Difference	=	17.72503	← 11
Pooled SD	=	46.89602	

Degrees of Freedom = 26 ← 12

Exit Command: <CR>=SP, R(erun), S(ystem), or PRUGRAM --

hipótesis $H_0: \mu_1 - \mu_2 = 0$,
 $H_a: \mu_1 \neq \mu_2$

- (4) El programa pide la definición de lo que el llama "arreglos x e y", esto es, la identificación de los dos grupos de datos a analizar, estos grupos pueden estar en el mismo (5) o diferente archivo y pueden tener igual o diferente número de observaciones
- (5) En este caso los dos conjuntos de valores están en el mismo archivo (TINDEP.DAT)
- (6) El "arreglo x" se define a ser la primera columna de TINDEP.DAT, el "Arreglo y" la segunda columna del mismo archivo. Note que en los cálculos se hará $t_c = (x - y)/s_d$. Así que el signo de t_c dependerá de la definición de los dos arreglos
- (7) Los dos conjuntos tienen igual número de observaciones (14). Note que si tuvieran tamaños diferentes de muestra "no podrían estar en el mismo archivo" pues STATPAK requiere archivos completos
- (8) $t_c = -2.7241$ ($\bar{x} < \bar{y}$)
- (9) Niveles de significancia para hipótesis alternas de una y dos colas
- (10) Descripción de las muestras (media y desviación estándar)
- (11) Valores usados para los cálculos

Resultados

- (i) diferencia de las medias
- (ii) error estándar de la diferencia de las medias. Este valor se puede usar para construir intervalos de confianza
- (iii) Desviación estándar combinada. En este caso es la raíz cuadrada del promedio simple de las varianzas puesto que $n_x = n_y$
- (12) Grados de libertad $(n_x - 1) + (n_y - 1) = 26$

4.4.1.3.3. Hipótesis sobre la media de una población

El cuadro 4.54 presenta un conjunto de 50 datos obtenidos con el programa RANUM y que se supone pertenecen a una población normal, con media $\mu = 1.74$ y varianza $\rho^2 = .0081$ (ver sección). En los cuadros 4.55.a y 4.55.b se prueba la hipótesis $H_0: \mu = 1.74$ bajo el supuesto de que la varianza poblacional es conocida o no respectivamente. Los puntos a notar se marcan con flechas numeradas.

- (1) Título del programa
- (2) Se quiere en ambos casos probar la hipótesis nula $H_0: \mu = 1.74$
- (3) En el primer caso (cuadro 4.55.a) se supone conocida la varianza poblacional y se indica su valor (.0081). En el segundo caso (cuadro 4.55.b) se indica que no se conoce la varianza
- (4) El archivo sobre el cual se trabaja es NUMALEAT.DAT
- (5) Descripción de la muestra. Note lo siguiente:
 - (i) en ambos casos $\bar{y} = 1.719618$

Cuadro 4.54

--> PROG: SPELD FILE: NUHALLAT.DAT DATE: 03-01-1986 ID: FALL 1

1: 1.645535
2: 1.671870
3: 1.675365
4: 1.566812
5: 1.64595
6: 1.790856
7: 1.721449
8: 1.772628
9: 1.757054
10: 1.749576
11: 1.729744
12: 1.804951
13: 1.715823
14: 1.565904
15: 1.849705
16: 1.68931
17: 1.70722
18: 1.704755
19: 1.70028
20: 1.680053
21: 1.627336
22: 1.780796
23: 1.655022
24: 1.772645
25: 1.74563
26: 1.756732
27: 1.73851
28: 1.609118
29: 1.72139
30: 1.693047
31: 1.66144
32: 1.647227
33: 1.523296
34: 1.813678
35: 1.751493
36: 1.661255
37: 1.75744
38: 1.674242
39: 1.635847
40: 1.783966
41: 1.778037
42: 1.684165
43: 1.893510
44: 1.807508
45: 1.800199
46: 1.745151
47: 1.768963
48: 1.787507
49: 1.54571
50: 1.872229

Cuadro 4.55 a

---> Single Sample t-statistic <---

Hypothesized Population Mean: ? 1.74 $\leftarrow 2$
 Is the Population Variance known (Y/N)? Y $\leftarrow 3$
 Population Variance: ? .0081 $\leftarrow 3$
 Input File (Q=Quit, <CR>=WORK.DAT): NUMALEAT.DAT $\leftarrow 4$
 1 Data Field(s)
 1 File Columns in NUMALEAT.DAT

Reading data.....

Mean = 1.719618
 Standard Deviation = 9.000001E-02 $\leftarrow 5$
 Standard Error = 1.272792E-02
 t-statistic = -1.601365 $\leftarrow 6$

Degrees of Freedom = * USE NORMAL TABLE * $\leftarrow 7$

One-sided Significance Level = .0546 $\leftarrow 8$
 Two-sided Significance Level = .1093

Exit Command: <CR>=SP, R(erun), S(ystem), or PROGRAM -- R

Cuadro 4.55 b

---> Single Sample t-statistic <---

Hypothesized Population Mean: ? 1.74 $\leftarrow 2$
 Is the Population Variance known (Y/N)? N $\leftarrow 3$
 Input File (Q=Quit, <CR>=NUMALEAT.DAT):
 1 Data Field(s) $\leftarrow 4$
 1 File Columns in NUMALEAT.DAT

Reading data.....

Mean = 1.717818
 Standard Deviation = 7.771311E-02 $\leftarrow 5$
 Standard Error = 1.077027E-02
 t-statistic = -1.00455 $\leftarrow 6$

Degrees of Freedom = 4 $\leftarrow 7$

One-sided Significance Level = .0548 $\leftarrow 8$
 Two-sided Significance Level = .0697

Exit Command: <CR>=SP, R(erun), S(ystem), or PROGRAM -- R $\leftarrow 9$

- (ii) cuando se asume varianza conocida el programa toma su valor para calcular la desviación estándar y el error estándar
 $(s = \sqrt{.0081} = .09, s_{\bar{y}} = s/\sqrt{50} = .0127)$. Cuando se desconoce la varianza poblacional el programa estima de la muestra los valores $(s = .0777, s_{\bar{y}} = .0109)$
- (6) Los valores de t_c difieren pero ambos son negativos puesto que $\bar{y} < 1.74$
- (7) Los grados de libertad se utilizan cuando se desconoce la varianza poblacional (cuadro 4.55.b), cuando se conoce la varianza $t_c \hat{=} z_c$ y basta usar las tablas de la distribución normal estándar (cuadro 4.55.a)
- (8) Es importante notar como el hecho de asumir que se conoce la varianza poblacional produce niveles de significación diferentes a aquellos obtenidos cuando no se hace la suposición

4.4.1.3.4. Comparación de dos poblaciones, muestras pareadas

El cuadro 4.56 muestra el archivo utilizado para este ejemplo, el cual corresponde a los datos presentados en el libro de Steel and Torrie, página 79, y que se refieren al efecto de dos presiones de vapor (4.4 mm Hg y 9.9 mm Hg) sobre concentración de azúcar en el néctar de claveles. Cada flor se partió por la mitad y cada mitad se sometió a una de las dos presiones, de esta forma cada observación corresponde a un mismo individuo tratado de dos formas diferentes. Se puede ver esta disposición experimental como un diseño de bloques al azar con 10 bloques (flores) y 2 tratamientos (presiones).

El cuadro 4.53 presenta las instrucciones y resultados de la prueba de hipótesis $H_0: \mu_1 = \mu_2$ donde μ_1 es la media poblacional de las flores sometidas a una presión

de 4.4. mm de mercurio, μ_2 la media poblacional para el otro tratamiento. Con flechas numeradas se señalan los siguientes puntos:

- (1) Título del programa
- (2) Ante la alternativa de probar una hipótesis nula general $H_0: \mu_1 - \mu_2 = k$, con k una constante, se elige $k = 0$, esto es $H_0: \mu_1 = \mu_2$
- (3) El análisis se hace sobre el archivo TPAR.DAT
- (4) Descripción de la prueba
 - (i) promedio de las diferencias (\bar{d})
 - (ii) desviación estándar de las diferencias
 - (iii) Error estándar de la media de las diferencias ($s_{\bar{d}}$)
- (5) Grados de libertad $(n-1)$ $(t-1) = (10-1)(2-1) = 9$, note que son los grados de libertad correspondientes al error en un diseño de bloques
- (6) Valor de $t_c = \bar{d}/s_{\bar{d}}$
- (7) Se rechaza H_0 con seguridad de "casi" 100% de no cometer error tipo I al hacerlo

Resultados

Cuadro 4.56

---> PROG: SPEED FILE: TPAR.DAT DATE: 03-01-1985 ID:

1: 62.5 51.7
 2: 65.2 54.2
 3: 57.6 53.3
 4: 69.9 57.0
 5: 69.4 56.4
 6: 70.1 61.5
 7: 67.8 57.2
 8: 67.0 56.2
 9: 78.5 58.4
 10: 62.4 55.8

Cuadro 4.57

---> Paired Samples t-Statistic --->

Hypothesized mean difference (CUR=0):

Input File (Q-Quit, C=Coefficient):

2 Data Field(s)

2 File Columns in INK.DAT

Field Assignment(s) (CUR=1):

Reading data.....

Mean Difference = 10.87

Standard Deviation = 3.23036

Standard Error = .7030753

Degrees of Freedom = 9

t-statistic = 15.45845

One-sided Significance Level = 2e-0000

Two-sided Significance Level = 2e-0000

Exit Command: <CR>=SP, R(=run), S(system), or H(=end)

4.5 Programas para análisis de Regresión y Correlación

4.5.1. CORREL

Covarianza y Correlación Lineal Simple

4.5.1.1. Descripción

Calcula la Covarianza y Correlación de cualquier número de pares de variables junto con sus estadísticos descriptivos, como media y desviación estándar. Calcula además el valor t - calculado para probar la hipótesis nula $H_0: \rho = 0$.

4.5.1.2. Funciones

Permite el análisis de asociación entre variables y una prueba de hipótesis para esta asociación. Si se usa junto con la opción "C" del programa MTESTS la prueba de hipótesis se completa con el cálculo del error tipo I estimado, aportado por este último.

4.5.1.3. Funcionamiento

Los pasos en la utilización de los dos programas son:

- (1) Teclee CORREL <CR> en la primera instrucción de STATPAK
- (2) Indique el número de variables (Columnas en el archivo) que entran al análisis
- (3) Responda a las instrucciones de definición de archivo y asignación de campos
- (4) Indique la forma de salida que desea (usualmente impresión)
- (5) En la instrucción de salida del programa (Exit Command) indique: MTEST
Exit Command CR = SP, R(erun), S(ystem) o PROGRAM MTESTS
- (6) STATPAK se dirigirá al Programa MTEST y presentará el menú de opciones de este programa. Teclee C <CR>
- (7) Informe al programa el número de observaciones, el coeficiente de correlación estimado por CORREL y el valor K_1 contra el que quiere hacer la prueba de hipótesis $H_0: \rho = k_1$

- (8) El programa calcula t_c para $H_0: \rho = 0$, intervalos de confianza para el verdadero valor de ρ y Z_c para $H_0: \rho = k$

4.5.1.4. Ejemplo

En la Sección 4.1.3. se introdujo un ejemplo sacado del libro de Steel y Torrie y que corresponde a 30 observaciones efectuadas en hojas de tabaco para estudiar la relación entre los contenidos de algunos componentes químicos (nitrógeno, cloro, potasio) y el tiempo de quemado de la hoja (en segs) expresado en escala logarítmica. Con el fin de medir el grado de asociación lineal de cada componente con el tiempo de quemado y las asociaciones entre componentes, que pueden indicar problemas de multicolinealidad en un futuro modelo de regresión múltiple, se hace un análisis de correlación de las cuatro variables, todas contra todas. El cuadro 4.58 presenta los pasos seguidos y los resultados obtenidos, como de costumbre se señalan con flechas numeradas algunos puntos de interés:

- (1) Se requieren correlaciones entre las cuatro variables
- (2) El archivo de entrada es TABACO.DAT
- (3) Salida por impresora
- (4) El usuario define el título que llevará cada hoja impresa
- (5) Se indica salir del programa CORREL y entrar al MTESTS con el fin de hacer pruebas de hipótesis e intervalos de confianza para algunas correlaciones
- (6) La salida del programa muestra la media y desviación estándar de cada variable (identificada bajo la letra A), su covarianza y correlación con cada una de las otras (identificada bajo la letra B) y el valor t-calculado para la hipótesis nula $H_0: \rho_{x_i x_j} = 0$
- (7) Número de observaciones procesadas
- (8) Ya en el programa MTESTS se pide la opción "C"
- (9) Se define el número de observaciones
- (10) El coeficiente de correlación estimado (ver No. 10 en anterior hoja) con el

Cuadro 4.58

----> CORRELATION & COVARIANCE ----

How many Data Fields? 4 ←

Input File (Q=Quit, <CR>=WORK.DAT) TABACO.DAT ←

4 Data Field(s)

4 File Columns in TABACO.DAT

Field Assignment(s) (<CR>=1:1):

Working.....

Output : S(creen), P(rinter), F(ile), or Q(uit)? P ←

--> MAKE PRINTER READY

Report ID: CALCULO DE CORRELACIONES ←

Output : S(creen), P(rinter), F(ile), or Q(uit)? Q ←

Exit Command: <CR>=BP, R(eturn), S(ystem), or PROGRAM -- MTESTS ←

--> PROG:CORREL FILE: TABACO.DAT DATE:05-20-1986 ID: CALCULO DE CORRELACIONES PAGE 1

A	MEAN	STD DEV	B	COVARIANCE	CORRELATION	T-TEST
1	3.278667	.5826677	2	7.383815E-02	.2094004	1.133165
			3	5.760503E-02	9.254283E-02	.4919084
2	.8076668	.605177	4	-.2003536	-.7177293	-5.454179
			3	.263233	.4073991	2.360529
3	4.653666	1.068078	4	-.1452235	-.4996385	-3.052107
4	.686	.4802844	4	9.20176E-02	.1793668	.964766

OF ROWS = 30 ←

cual se quieren hacer la prueba y los intervalos.

- (11) La opción "C" de MTESTS permite probar la hipótesis nula $H_0: \rho = k$ con k cualquier valor entre -1 y 1 , en este caso $K = 0$
- (12) MTESTS muestra el valor t -calculado para el caso en que $K = 0$ (observe el No. 12 en la hora anterior), siempre para ese valor de K
- (13) Los errores tipo I estimados, que en este caso son $.0000$ lo cual indica que se puede rechazar la hipótesis nula con "casi" 100% de confianza y aceptar cualquiera de las alternas: $H_a: \rho < 0$ o $H_a: \rho > 0$ (pruebas de una cola/One-Sided) o $H_a: \rho \neq 0$ (prueba de dos colas/Two-Sided)
- (14) Intervalo de confianza para el verdadero valor poblacional. En este caso indica que: la probabilidad de que el verdadero grado de asociación entre las variables 1 (% de nitrógeno) y 4 (Log. del tiempo de quemado) esté comprendido entre los valores $-.8565$ y $-.4821$ es 95%
- (15) Las tres líneas señaladas corresponden a la prueba de hipótesis $H_0: \rho = K$ cuando $K = 0$. En este caso no es de interés
- (16) Se indica que no se sale de MTEST opción "C" sino que se va a trabajar con otro valor de correlación
- (17) Definido el No. de observaciones se indica el nuevo coeficiente de correlación de interés, en este caso ρ_{x_3y} (potasio vs tiempo de quemado, ver N^o. 17 en la hoja anterior)
- (18) MTESTS muestra el valor de t calculado (ver No. 18 de la anterior hoja)
- (19) En este caso las probabilidades de rechazar por error la hipótesis nula $H_0: \rho = 0$ son 17.13% en caso de que la

hipótesis alterna sea $\rho < 0$ o $\rho > 0$ y 34.26%, el doble, si la hipótesis alterna es $\rho \neq 0$

- (20) Intervalo de 95% de confianza para el verdadero valor ρ x_4y
- (21) Prueba de hipótesis $H_0: \rho = K$, en este caso con $K = 0$. Note la similitud de los niveles de error en esta prueba y la de t , numeral 19

--- MEANS TESTS ---

C orrelation Coefficient t-statistic
 O rdinary t-statistic, Unpaired Samples
 P earson Product Moment Correlation
 S ingle Sample t-statistic
 t -statistic, Paired Samples

X (e)xit this program)

Select desired function (C,O,P,S,T,X)? C

---> Correlation Coefficient t-statistic <---

of data items (N)? 30

Estimate of Correlation Coefficient (r)? -.7177293

Hypothesized value for $\rho(0)$ (CR)=0?t-statistic for $\rho=0$: -5.454179

One-sided Significance Level = .0000

Two-sided Significance Level = .0000 .

Approximate 95% C.I. [-0.8565 to -0.4821]

z-statistic for hypothesis: -4.691845

One-sided Significance Level = .0000

Two-sided Significance Level = .0000

Exit Command: (CR)=SF, R(erun), S(ystem), or PROGRAM -- R

---> Correlation Coefficient t-statistic <---

of data items (N)? 30

Estimate of Correlation Coefficient (r)? .1793668

Hypothesized value for $\rho(0)$ (CR)=0?t-statistic for $\rho=0$: .9647661

One-sided Significance Level = .1713

Two-sided Significance Level = .3426

Approximate 95% C.I. [-0.1934 to 0.5069]

z-statistic for hypothesis: .9422098

One-sided Significance Level = .1730

Two-sided Significance Level = .3461

Exit Command: (CR)=SF, R(erun), S(ystem), or PROGRAM --

4.5.2. MLINREG

Análisis de regresión múltiple

4.5.2.1. Descripción

MLINREG es uno de los programas más potentes de STATPAK. Permite hacer análisis de regresión múltiple a un conjunto de datos con cualquier número de observaciones y hasta 50 variables, en computadoras con 128 K de memoria RAM o 25 en aquellas con 64 K. Las características de la microcomputación hacen que los tiempos para análisis de conjuntos grandes de datos sean largos, por ejemplo, una regresión con 400 observaciones y 10 variables independientes puede demorar aproximadamente 15 minutos.

4.5.2.2. Funciones

Con MLINREG es posible enfrentar el problema de regresión de cuatro formas diferentes.

- (1) Regular: produce un análisis clásico, incluyendo estimaciones y prueba de hipótesis, para un único modelo indicado por el usuario.
- (2) Selección hacia adelante (FORWARD REGRESSION): construye un modelo que satisface determinadas condiciones, dadas por el usuario, respecto a las variables incluidas. Forward comienza con un modelo que contiene sólo el término de intercepto y va añadiendo variables, una por una, siempre y cuando su importancia, medida en términos de F parcial, sea significativa.
- (3) Selección hacia atrás (BACKWARD REGRESSION): construye un modelo con iguales características que forward pero en este caso inicia con uno que contiene todas las posibles variables y va sacando, una por una, aquellas que no satisfacen determinado nivel de significancia (F parcial) definido por el usuario.
- (4) Selección interactiva: esta opción es la preferida de quien escribe, por el manejo conceptual que implica en el usuario su correcta utilización. La opción comienza incluyendo en el modelo el término de intercepto y presenta al usuario las F parciales de él y de los

términos no incluidos. El usuario selecciona, paso por paso, qué variables independientes entran o salen del modelo hasta lograr aquel que lo satisface más dentro de los posibles. Por razones didácticas se trabajará a fondo esta opción en el manual, aunque se dan ejemplos de las otras opciones.

Independientemente de cual tipo de manejo se escoja el programa ofrece la posibilidad de hacer el cálculo de los residuales y valores predichos formando un archivo de datos que los incluye y una prueba de Durbin Watson para el análisis de autocorrelación de primer orden entre errores residuales.

4.5.2.3. Funcionamiento

Los pasos necesarios para la utilización del programa son los siguientes:

- (1) Teclee MLINREG <CR> en la primera instrucción STATPAK
- (2) Seleccione el tipo de trabajo de regresión que desea (regular, forward, backward, interactiva)
- (3) Si desea un archivo que contenga valores "y" observados y esperados junto con los residuales, indique a STATPAK el nombre que debe darle a ese archivo. Si no se define un nombre no se calculan los valores. Si no los va a usar posteriormente omita su cálculo lo cual le dará un considerable ahorro de tiempo.
- (4) Indique si desea o no (Y/N) la prueba de Durbin Watson.
- (5) Si escogió la opción backward o forward defina al programa un valor F crítico tal que si el valor F parcial para una variable es mayor que él, la variable sea incluida en el modelo (Forward) o, si el valor F parcial es menor que él, la variable sea eliminada (Backward) del modelo. En este punto conviene notar lo siguiente: suponga que tiene n observaciones y p variables

independientes. Los grados de libertad del error en el modelo que incluye todas las variables son $(n-p)$, así, se puede escoger como F crítico, conservadoramente, un valor de tablas $F(1, n-p, \alpha)$, donde α es el nivel de significancia deseado.

- (6) Defina el número de variables independientes que entrarán al análisis.
- (7) Ejecute los pasos "Input File" y "Field assignment" teniendo presente que la primera variable asignada será tomada como variable dependiente.
- (8) En este punto la máquina indicará que está leyendo los datos, recuerde que este paso es el más largo, tenga paciencia. La computadora mostrará un punto en pantalla por cada observación leída y procesada.
- (9) En caso de que la opción de trabajo fuera regular, Forward o Backward el programa hace el análisis y muestra al usuario la instrucción de salida, en caso de regresión interactiva se continúa con los siguientes pasos:
- (10) STATPAK presenta al usuario en la pantalla un listado con los números de las variables en análisis indicando cuáles están en el modelo (la primera vez solo el intercepto, β_0) y cuales no, junto con los valores F parciales para cada una de ellas y el coeficiente de determinación R^2 para el modelo en pantalla. El usuario debe seleccionar qué variable entra a la ecuación o sale de ella, tecleando su número de identificación (orden que le correspondió en la asignación de campos) y <CR> .
- (11) Cuando el modelo construido satisfaga al usuario este lo indica pidiendo una salida de resultados e indicando a la máquina qué tipo de salida desea.

4.5.2.4. Ejemplos

En lo que sigue se da un grupo de ejemplos y su respectiva discusión, para mostrar con el mismo conjunto de datos el funcionamiento de las diferentes opciones. Los datos corresponden al experimento con tabaco introducido en la Sección 4.1.3. y que aparece listado en el cuadro 4.14. Se tienen 30 observaciones (hileras) y cuatro variables (columnas) que corresponden a porcentaje de nitrógeno (x_1), de cloro (x_2) y de potasio (x_3) y el logaritmo natural del tiempo de quemado de la hoja (y).

Ejemplo (1): Opción regular: Los cuadros 4.59.a y 4.59.b presentan el análisis clásico de regresión en el cual no se hace ningún tipo de selección de variables con flechas numeradas se señalan las siguientes observaciones:

- (1) Selección de la opción regular
- (2) Se indica que calcule residuales y valores "Y" estimados y los grave en un archivo con el nombre TABACO.RSD. El programa advierte que ya hay un archivo con ese nombre (ver cuadro) y pregunta si escribe encima de él, a lo cual se le responde que sí.
- (3) Se ordena que no calcule la Durbin-Watson.
- (4) Entran a análisis tres variables independientes.
- (5) Definición del archivo de entrada y asignación de campos.
- (6) STATPAK anuncia qué está leyendo (y procesando simultáneamente) los datos y va escribiendo en pantalla un punto por observación leída y procesada.
- (7) Terminó los cálculos iniciales, pregunta el tipo de salida requerido, Se le indica la impresora.
- (8) Los resultados del análisis inician con una línea de identificación en donde se informa el nombre del programa y del archivo sobre el cuál operó, la fecha y el título, puesto por el usuario al pedir salida por impresora.
- (9) El primer bloque de resultados corresponde a los estimadores de los coeficientes de regresión, contiene:

una identificación que corresponde a la asignación de campos de las variables el valor del coeficiente, su desviación, el valor t calculado para probar la hipótesis " $H_0: \beta_1 = 0$ dado que todas las demás variables ya están en el modelo" (t-parcial), la correlación parcial de cada variable con la dependiente y la contribución "parcial" al R^2 (cuanto aporta al R^2 cuando entra última al modelo).

- (10) El segundo bloque de resultados muestra el análisis de varianza para el modelo. Presenta fuentes de variación, sumas de cuadrados, grados de libertad y cuadrados medios. Las fuentes de variación son "modelo" (todas las variables incluidas), "Error Residual", y "total".
- (11) El último bloque presenta los valores R^2 , F calculado para H_0 : el modelo no explica la variación en "y", R^2 corregido (!) y el error estándar de la regresión ($\sqrt{CM_{error}}$)

Ejemplo (2). Opción FORWARD: en este caso se pide una selección de variables "Hacia Adelante". El cuadro 4.60 muestra la aplicación de esta opción a los datos anteriores. Con flechas numeradas se señalan los siguientes puntos:

- (1) Se escoge la opción Forward.
- (2) No se requiere cálculo de residuales
- (3) F crítico para la selección de variables. En este ejemplo se dio $F(1,27,.05) = 4.21$. Los grados de libertad totales, son 29 y entonces los grados de libertad del error pueden ser: 28 si solo entra una variable al modelo, 27 si dos y 26 si tres, así los valores de F crítico para la prueba de la significancia de una variable serían:

$F(1, 29, .05) = 4.18$ si no hay ninguna variable en el modelo (primera selección)

Cuadro 4.59 a

----> MULTIPLE LINEAR REGRESSION <----
Select regression: Regular, Forward, Backward, Interactive (R,F,B,I)? R

Residual File Name (<CR> to skip): TABACO.RSD
TABACO.RSD exists: re-use it (Y,N)? Y
Calculate Durbin-Watson (Y/N)? N

of Independent Variables (P)? 3

Field 1 = Dependent Variable (Y)
Fields 2 - 4 = Independent Variable (X) Matrix


Input File (Q=Quit, <CR>=TABACO.DAT):
4 Data Field(s)
4 File Columns in TABACO.DAT
Field Assignment(s) (<CR>=1:1): 4 1 2 3

Reading Data.....

Output : S(creen), P(rinter), or Q(uit)? P



Cuadro 4.59 b

 --> PROB:MLINREG FILE: TABACO.DAT DATE: 04-03-1986 ID: REGRESSION MULTIPLE, OPCION REGULAR
PAGE 1

TERM	COEFFICIENT	STD. ERROR	T-STATISTIC	PART. CORR	CONTR. R-SQ
B 0	1.811043	.2795193	6.479133	---	---
B 1	-.5314553	6.957676E-02	-7.638401	0.6917	0.3974
B 2	-.4396359	7.303726E-02	-6.019337	0.5822	0.2468
B 3	.2089753	4.064021E-02	5.142082	0.5042	0.1801

143



	SUM SQ	DEG FR	MEAN SQ
DUE TO REGRESSION	5.504734	3	1.834911
ABOUT REGRESSION	1.184785	26	4.556867E-02
TOTAL	6.68952	29	.2306731




R-SQUARED: .8228893
F-TEST: 40.26695

CORRECTED R-SQUARED: .8024535
STD ERROR OF REG: .2134682



Cuadro 4.60

----> MULTIPLE LINEAR REGRESSION <----

Select regression: Regular, Forward, Backward, Interactive (R,F,B,I)? F 

Residual File Name (<CR> to skip): 

Calculate Durbin-Watson (Y/N)? N

F to Enter Cutoff? 4.21 

of Independent Variables (P)? 3

Field 1 = Dependent Variable (Y)

Fields 2 - 4 = Independent Variable (X) Matrix

Input File (Q=Quit, <CR>=TABACO.DAT):

4 Data Field(s)

4 File Columns in TABACO.DAT

Field Assignment(s) (<CR>=1:1): 4 1 2 3

Reading Data.....

Output : S(screen), P(printer), or Q(uit)? P

Cuadro 4.60 (continuación)

--> PROG:MLINREG FILE: TABACO.DAT DATE: 04-03-1986 ID: REGRESION MULTIPLE, OPCION FORWARD PAGE 1

TERM F TO ENTER
 B 1 29.74806
 B 2 9.315353
 B 3 .9307736

F to Enter Cutoff: 4.21

Adding Variable # 1

--> PROG:MLINREG FILE: TABACO.DAT DATE: 04-03-1986 ID: REGRESION MULTIPLE, OPCION FORWARD PAGE 2

TERM	COEFFICIENT	STD. ERROR	T-STATISTIC	PART. CORR	CONTR. R-SQ
B 0	2.625704	.3610242	7.27293		
B 1	-.5916136	.1084698	-5.454179	0.5151	0.5151

	SUM SQ	DEG FR	MEAN SQ
DUE TO REGRESSION	3.446007	1	3.446007
ABOUT REGRESSION	3.243512	28	.1158397
TOTAL	6.68952	29	.2306731

R-SQUARED: .5151352 CORRECTED R-SQUARED: .4978186
 F-TEST: 29.74806 STD ERROR OF REG: .3403524

Cuadro 4.60 (continuación)

--> PROG:MLINREG FILE: TABACO.DAT DATE: 04-03-1986 ID: REGRESION MULTIPLE, OPCION FORWARD PAGE 3

TERM F TO ENTER

 B 2 9.647292
 B 3 3.881351

F to Enter Cutoff: 4.21

Adding Variable # 2

--> PROG:MLINREG FILE: TABACO.DAT DATE: 04-03-1986 ID: REGRESION MULTIPLE, OPCION FORWARD PAGE 4

TERM	COEFFICIENT	STD. ERROR	T-STATISTIC.	PART. CORR	CONTR. R-SQ
B 0	2.653132	.3156926	8.404161	---	---
B 1	-.5285493	9.696244E-02	-5.451073	0.5239	0.3931
B 2	-.2899645	9.335597E-02	-3.106009	0.2632	0.1276

	SUM SQ	DEG FR	MEAN SQ
DUE TO REGRESSION	4.299853	2	2.149926
ABOUT REGRESSION	2.389667	27	8.850619E-02
TOTAL	6.68952	29	.2306731

R-SQUARED: .6427745 CORRECTED R-SQUARED: .6163133
 F-TEST: 24.29126 STD ERROR OF REG: .2974999

Cuadro 4.60 (continuación)

--> PROG:MLINREG FILE: TABACO.DAT DATE: 04-03-1986 ID: REGRESION MULTIPLE, OPCION FORWARD PAGE 5

TERM F TO ENTER

 B 3 26.44101

F to Enter Cutoff: 4.21

Adding Variable # 3

--> PROG:MLINREG FILE: TABACO.DAT DATE: 04-03-1986 ID: REGRESION MULTIPLE, OPCION FORWARD PAGE 6

TERM	COEFFICIENT	STD. ERROR	T-STATISTIC	PART. CORR	CONTR. R-SQ
---	---	---	---	---	---
B 0	1.811043	.2795193	6.479133	---	---
B 1	-.5314553	6.957676E-02	-7.638401	0.6917	0.3974
B 2	-.4396359	7.303726E-02	-6.019337	0.5822	0.2468
B 3	.2089753	4.064021E-02	5.142082	0.5042	0.1801

	SUM SQ	DEG FR	MEAN SQ
---	---	---	---
DUE TO REGRESSION	5.504734	3	1.834911
ABOUT REGRESSION	1.184785	26	4.556867E-02
TOTAL	6.68952	29	.2306731

R-SQUARED: .8228893 CORRECTED R-SQUARED: .8024535
 F-TEST: 40.26695 STD ERROR OF REG: .2134682

Cuadro 4.60 (continuación)

--> PROG:MLINREG FILE:TABACO.DAT DATE:04-03-1986 ID:REGRESION MULTIPLE, OPCION FORWARD
PAGE 7

TERM F TO ENTER

F to Enter Cutoff: 4.21

--> No more variables to Enter

--> PROG:MLINREG FILE:TABACO.DAT DATE:04-03-1986 ID:REGRESION MULTIPLE, OPCION FORWARD
PAGE 8

STEP	F TO ENTER	TERM	R-SQR.	F-TEST
-----	-----	-----	-----	-----
1	29.748	B 1	0.5151	29.74806
2	9.647	B 2	0.6428	24.29126
3	26.441	B 3	0.8229	40.26695



$F(1, 28, .05) = 4.20$ si ya entró una (segunda selección), y

$F(1, 27, .05) = 4.21$ si entraron ya dos variables.

Así, el valor 4.21 es el mayor de los posibles "F-críticos" para la decisión de incluir o no una variable en cualquier etapa.

- (4) En la primera etapa de selección se debe escoger una de las tres variables, Forward selecciona la que presente mayor F parcial (en la discusión sobre la opción interactiva se explicará claramente este concepto)...
- (5) ... y cumpla con el requisito F_c $F(1, 27, .05)$...
- (6) En este caso la variable número 1
- (7) Luego presenta los resultados de esta primera etapa y pasa a la siguiente.
- (8) Al terminar Forward presenta un resumen del proceso, mostrando la variable incluida en cada paso, el valor de F_2 parcial con que entró al modelo y los R^2 y F calculados para el modelo total resultante en cada paso.

Ejemplo (3): Opción BACKWARD: de nuevo con los mismos datos, se muestra la forma de selección de variables hacia atrás. El cuadro 4.61 presenta los resultados con algunas puntualizaciones marcadas con flechas numeradas:

- (1) Selección Backward
- (2) Valor crítico de $F = 4.21$, en este caso toda variable que presente un F parcial menor que 4.21 será eliminada del modelo
- (3) STATPAK presenta los resultados de incluir todas las variables en el modelo
- (4) Con el fin de eliminar la primera variable se calculan los F parciales para todas ellas, en este caso ninguna es menor que 4.21 así que no hay eliminación


Cuadro 4.61

----> MULTIPLE LINEAR REGRESSION <----

Select regression: Regular, Forward, Backward, Interactive (R,F,B,I)? B 

Residual File Name (<CR> to skip):

Calculate Durbin-Watson (Y/N)? N

F to Remove Cutoff? 4.21 

of Independent Variables (P)? 3

Field 1 = Dependent Variable (Y)

Fields 2 - 4 = Independent Variable (X) Matrix

Input File (Q=Quit, <CR>=TABACO.DAT):

4 Data Field(s)

4 File Columns in TABACO.DAT

Field Assignment(s) (<CR>=1:1): 4 1 2 3

Reading Data.....

Output : S(screen), P(printer), or Q(uit)? P

Cuadro 4.61 (continuación)

3

--> PROG:MLINREG FILE:TABACO.DAT DATE:04-03-1986 ID:REGRESION MULTIPLE, OPCION BACKWARD
PAGE 1

TERM	COEFFICIENT	STD. ERROR	T-STATISTIC	PART. CORR	CONTR. R-SQ
B 0	1.811043	.2795193	6.479133	---	---
B 1	-.5314553	6.957676E-02	-7.638401	0.6917	0.3974
B 2	-.4396359	7.303726E-02	-6.019337	0.5822	0.2468
B 3	.2089753	4.064021E-02	5.142082	0.5042	0.1801

	SUM SQ	DEG FR	MEAN SQ
DUE TO REGRESSION	5.504734	3	1.834911
ABOUT REGRESSION	1.184785	26	4.556867E-02
TOTAL	6.68952	29	.2306731

R-SQUARED: .8228893 CORRECTED R-SQUARED: .8024535
F-TEST: 40.26695 STD ERROR OF REG: .2134682

Cuadro 4.61 (continuación)

--> PROG:MLINREG FILE:TABACO.DAT DATE:04-03-1986 ID:REGRESION MULTIPLE, OPCION BACKWARD PAGE 2

TERM	F TO REMOVE
B 1	58.34518
B 2	36.23242
B 3	26.44101



F to Remove Cutoff: 4.21

--> No more variables to Remove

--> PROG:MLINREG FILE:TABACO.DAT DATE:04-03-1986 ID:REGRESION MULTIPLE, OPCION BACKWARD PAGE 3

STEP	F TO REMOVE	TERM	R-SQR.	F-TEST
----	-----	----	-----	-----



- (5) Resumen final de la selección de variables realizada. En este caso no se hizo ninguna eliminación

Ejemplo (4) Opción interactiva: Con los mismos datos se discute la utilización de esta opción paso por paso, su estudio detallado va a permitir no solo la solución del problema práctico sino la comprensión de los otros dos métodos de selección. El cuadro 4.62 presenta los pasos ejecutados y los resultados obtenidos. Con flechas se señalan los siguientes puntos de interés:

- (1) Selección de la opción interactiva
- (2) No se requiere cálculo de valores esperados y residuales ni prueba de Durbin Watson
- (3) Definición del número de variables independientes
- (4) Asignación de campos. El primer campo asignado corresponde a la variable dependiente (y), en este caso la columna 4 del archivo, logaritmo del tiempo de quemado de la hoja. El segundo corresponde a x_1 , el tercero a x_2 y el cuarto a x_3 . Al hacer esta definición queda establecido que los términos B_1 , B_2 , B_3 , que se manejarán en lo que sigue, corresponden a nitrógeno, cloro y potasio, respectivamente.
- (5) El programa establece el modelo $Y = \beta_0 + \epsilon$ y presenta el valor F calculado para el término β_0 .
- (6) Se identifican los términos no incluidos y el valor F calculado parcial para cada uno de ellos en este primer modelo, esto es, la F que presentaría cada uno si entrara "solo él", en este momento, al modelo. De otra forma, la F calculada para las hipótesis $H_0: \beta_1 = 0$ dado el modelo Y

$= \beta_0 + \beta_1 x_1 + \epsilon$, $H_0: \beta_2 = 0$ dado el modelo $Y = \beta_0 + \beta_2 x_2 + \epsilon$ y $H_0: \beta_3 = 0$ en el modelo $Y = \beta_0 + \beta_3 x_3 + \epsilon$

- (7) El coeficiente de determinación para el modelo $Y = \beta_0 + \epsilon$ es -1.9×10^{-8} (0.0 en términos prácticos)
- (8) Se indica a STATPAK que incluya el término B_1 en el modelo, con la instrucción 1 <CR>, puesto que tiene el mayor valor de F.
- (9) El modelo ahora considerado es $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
- (10) Las F parciales para x_2 y x_3 son ahora $F(\beta_2/\beta_0, \beta_1) = 9.65$ y $F(\beta_3/\beta_0, \beta_1) = 3.88$ Note que ahora miden la importancia de incluir β_2 o β_3 en el modelo "que ya contiene β_0 y β_1 "
- (11) El coeficiente de determinación para el modelo $Y = \beta_0 + \beta_1 x_1 + \epsilon$ es 51.5%
- (12) Se ordena incluir el término B_2 al modelo
- (13) El modelo ahora considerado es $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$
- (14) Observe como se incrementó F para B_3 al incluir B_2 en el modelo: $F(\beta_3/\beta_0, \beta_1) = 3.88$, $F(\beta_3/\beta_0, \beta_1, \beta_2) = 26.44$

Este comportamiento está indicando que la inclusión de β_2 reduce la suma de

cuadrados del error sin reducir la de β_3 . De otra forma, indica que x_2 y x_3 explican partes diferentes de la variación de "Y" y esto es claro si se observa que la correlación $\gamma_{x_2x_3} = .1794$ es no significativa (vea la anterior sección, recordando que x_2 y x_3 son las columnas 3 y 4 del archivo respectivamente)

(15) Se pide a STATPAK incluir la variable x_3

(16) El modelo ahora es $Y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \epsilon$

(17) Con $R^2 = 82.29\%$

(18) Se acepta este modelo indicando a la máquina que pase a la sección Output (0 <CR>)

(19) Se indica que dé la salida por la impresora

(20) STATPAK presenta los tres bloques de información ya descritos anteriormente. Las conclusiones son entonces:

i. Modelo estimado

$$\hat{y} = 1.81 - .53x_1 - .44x_2 + .21x_3$$

ii. Desviaciones estándar de los estimadores

Cuadro 4.62 (continuación)

IN equation:	TERM	F TO REMOVE
	B 0	52.89551
	B 1	29.74807 ← 9

NOT in equation:	TERM	F TO ENTER
	B 2	9.647292 ← 8
	B 3	3.881351

R-SQUARED = .5151352 ← 11

TERM to Enter/Remove, O)utput, Q)uit, <CR>=F stats ? 2 ← 8

IN equations:	TERM	F TO REMOVE
	B 0	70.62993
	B 1	29.71419 ← 13
	B 2	9.647292

NOT in equation:	TERM	F TO ENTER
	B 3	26.44101 ← 9

R-SQUARED = .6427745

TERM to Enter/Remove, O)utput, Q)uit, <CR>=F stats ? 3 ← 5

IN equations:	TERM	F TO REMOVE
	B 0	41.97917
	B 1	58.34518
	B 2	36.23242 ← 16
	B 3	26.44101

NOT in equations:	TERM	F TO ENTER
-------------------	------	------------

R-SQUARED = .8228893 ← 11

TERM to Enter/Remove, O)utput, Q)uit, <CR>=F stats ? 0 ← 6

Output : S(creen), P(rinter), or Q(uit)? P ← 11

$$\hat{\beta}_1 = -.53 + .069$$

$$\hat{\beta}_2 = -.44 + .073$$

$$\hat{\beta}_3 = .21 + .041$$

iii. Prueba de hipótesis "parciales":

valores de t - calculado para H_0 :

$\beta_1 = 0$, dado que todos los demás términos están en el modelo. Note que si estos valores se elevan al cuadrado se obtienen los valores F del numeral 16. En este caso

$$t_c (\beta_1 / \beta_0, \beta_2, \beta_3) = -7.64$$

$$t_c (\beta_2 / \beta_0, \beta_1, \beta_3) = -6.02$$

$$t_c (\beta_3 / \beta_0, \beta_1, \beta_2) = 5.14$$

El valor crítico de tablas con $\alpha = .05$ y 26 grados de libertad del error residual es $t(.05, 26) = 2.056$, por lo cual se declara significativo el efecto de las tres variables teniendo en cuenta que los contenidos de nitrógeno y cloro reducen el tiempo de quemado mientras que el potasio lo incrementa

iv. Correlaciones parciales: son valores de correlación entre cada variable independiente y la dependiente, ajustadas "ambas" por las variables restantes. Esto es: $r_{yx_1, x_2, x_3} = .6917$ indica que para valores constantes de x_2 y x_3 las variables Y y X_1 tienen una correlación positiva de .69

v. Contribución parcial de cada variable al R^2 del modelo

Cuadro 4.62 (continuación)

--> PROG:MLINREG FILE:TABACO.DAT DATE:04-03-1986 ID:REGRESION MULTIPLE, OPCION INTERACTIVA
PAGE 1

TERM	COEFFICIENT	STD. ERROR	T-STATISTIC	PART. CORR	CONTR. R-SQ
B 0	1.811043	.2795193	6.479133	---	---
B 1	-.5314553	6.957676E-02	-7.638401	0.6917	0.3974
B 2	-.4396359	7.303726E-02	-6.019337	0.5822	0.2468
B 3	.2089753	4.064021E-02	5.142082	0.5042	0.1801



	SUM SQ	DEG FR	MEAN SQ
DUE TO REGRESSION	5.504734	3	1.834911
ABOUT REGRESSION	1.184785	26	4.556867E-02
TOTAL	6.68952	29	.2306731

R-SQUARED: .8228893
F-TEST: 40.26695
CORRECTED R-SQUARED: .8024535
STD ERROR OF REG: .2134682

vi. Análisis de varianza para la regresión en los términos usuales

vii. Coeficiente de determinación (R^2), F calculada para H_0 : no hay explicación de "y" con las variables X incluidas y error estándar de la regresión ($\sqrt{CMerror}$). En este caso $R^2 = 82.29\%$, $F_c = 40.27$ y $s = .2135$

4.5.3. POLYREG

Análisis de regresiones polinomiales

4.5.3.1. Descripción: POLYREG es un programa construido para analizar con la metodología de mínimos cuadrados, la relación de dependencia entre dos variables cuando se espera que esta relación pueda ser explicada con un polinomio de la forma

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_p x^p + \epsilon_t, \quad \text{con } p < n$$

(número de observaciones). La diferencia con MLINREG es que POLYREG genera los valores x^2, x^3, \dots, x^p , mientras que MLINREG precisaría que se le dieran estos valores.

4.5.3.2. Funciones: Muchas relaciones biológicas y económicas entre variables han sido explicadas con modelos polinomiales, la razón de esto es que un polinomio de grado p ajusta una curva con $p-1$ cambios de pendiente, aproximando muchos tipos de respuesta. POLYREG permite al usuario estudiar polinomios, que expliquen la respuesta, presente en sus datos, hasta de grado $p = n - 1$, siendo n el número de observaciones.

4.5.3.3. Funcionamiento: los pasos para la utilización del programa son los siguientes:

- (1) Teclee POLYREG CR en la primera instrucción STATPAK
- (2) El programa pregunta por el orden del polinomio, indíquelo
- (3) Defina si desea el cálculo de valores predichos y residuales dando un nombre para el archivo en que se grabarán, si no define el archivo no se hacen los cálculos
- (4) Indique si desea prueba de Durbin-Watson para autocorrelación de errores
- (5) Defina el archivo de entrada y la asignación de campos
- (6) Indique el tipo de salida que desea

4.5.3.4. Ejemplo: en el cuadro 4.63 se presenta un conjunto de datos simulados en la computadora utilizando el programa RANUM y una función logarítmica $Y_1 = 2 + 3 \log(x_1)$, sumando a cada

valor $2 + 3 \log(x_1)$ un número aleatorio de una distribución normal $(0, .25)$ generado con RANUM. La figura 4.64 muestra la gráfica de x vs. y en donde se puede apreciar la curvatura de los datos. Con el fin de ejemplificar el uso de POLYREG se ajusta un modelo cuadrático $Y = \beta_0 +$

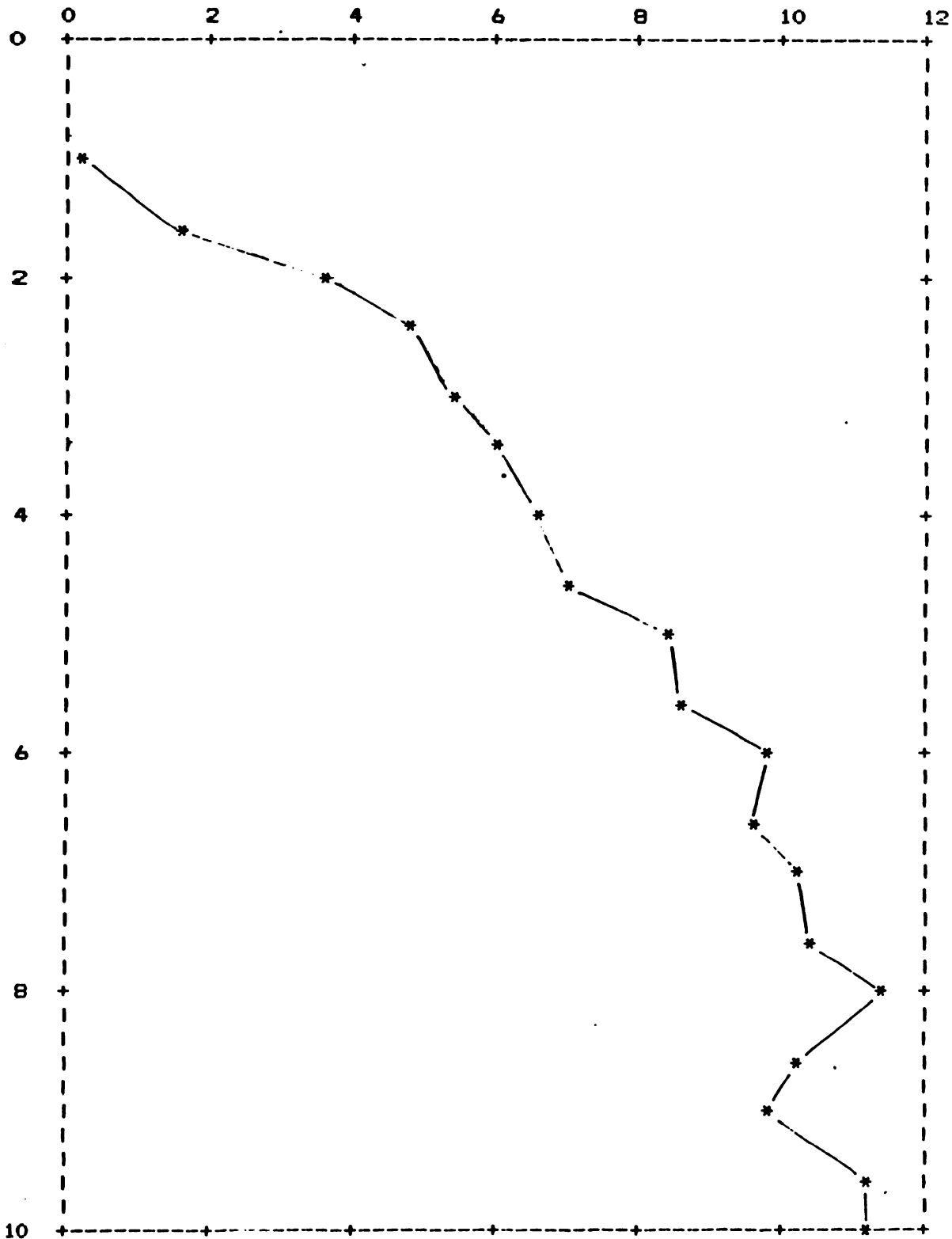
Cuadro 4.63

--> PROG: SPEED FILE: ONEV.DAT DATE: 06-24-1986 ID:

	X	Y
1:	1.0	.1400209
2:	1.5	1.660309
3:	2.0	3.635089
4:	2.5	4.871137
5:	3.0	5.373233
6:	3.5	6.009323
7:	4.0	6.514935
8:	4.5	6.978947
9:	5.0	8.456581
10:	5.5	8.686319
11:	6.0	9.880029
12:	6.5	9.506448
13:	7.0	10.16076
14:	7.5	10.43481
15:	8.0	11.32288
16:	8.5	10.23935
17:	9.0	9.700434
18:	9.5	11.14074
19:	10.	11.25653
20:	10.5	11.89964

X * 1 Y * 1 COLLISION MARK: ?

FILE: ONEV.DAT Columns: X = 1 (*) = 2



$\beta_1 x + \beta_2 x^2 + \epsilon$ a los datos. El cuadro 4.65 muestra las instrucciones utilizadas y el 4.66 los resultados del análisis. Con flechas numeradas se indican los siguientes puntos:

- (1) El orden del polinomio (cuadrático)
- (2) Se indica que no calcule residuales ni prueba de de DURBIN-WATSON
- (3) El archivo para analizar es ONEV.DAT
- (4) Salida por impresora
- (5) Al pedir la impresión de los resultados STATPAK da oportunidad de incluir un título que encabezará las hojas de resultados
- (6) El modelo estimado es $Y = -1.63 + 2.64 X - .136 X^2$
- (7) Valores de t para las pruebas de hipótesis sobre los coeficientes de regresión. En la discusión del programa ONEVREG se hará una discusión sobre estos valores
- (8) $R^2 = 92.27 \%$ y $F_c = 303.08 **$ para el modelo. Como conclusión inicial se puede indicar que el modelo cuadrático ha sido eficiente para explicar el comportamiento de los datos

Cuadro 4.65

```

      ---> POLYNOMIAL REGRESSION <---

Power of Polynomial (P)? 2 <1
Residual File Name (<CR> to skip):
Calculate Durbin-Watson (Y/N)? N <2
Field 1 = Independent Variable (X)
Field 2 = Dependent Variable (Y)

Input File (Q=Quit, <CR>=WORK.DAT): ONEV.DAT <3
  2 Data Field(s)
  2 File Columns in ONEV.DAT
Field Assignment(s) (<CR>=1:1):

Reading Data.....

Output : S(creen), P(rinter), or Q(uit)? P <4

--> MAKE PRINTER READY <5
Report ID: EJEMPLO, USO DEL PROGRAMA POLYREG.

```

Cuadro 4.66

--> PROG: POLYREG FILE: ONEV.DAT DATE: 06-24-1986 ID: EJEMPLO, USO DEL PROGRAMA POLYREG. PAGE 1

TERM	COEFFICIENT	STD. ERROR	T-STATISTIC	PART. CORR	CONTR. R-SQ
B 0	-1.634289	.5294269	-3.086903	---	---
B 1	2.639704	.2096288	12.59228	0.9032	0.2545
B 2	-.1365712	1.779009E-02	-7.676814	0.7761	0.0946

	SUM SQ	DEG FR	MEAN SQ
DUE TO REGRESSION	210.4977	2	105.2488
ABOUT REGRESSION	5.903513	17	.3472655
TOTAL	216.4012	19	11.38954

R-SQUARED: .9727196
 F-TEST: 303.0789
 CORRECTED R-SQUARED: .9695101
 STD ERROR OF REG: .5892924

4.5.4. ONEVREG

Regresiones con una variable independiente

4.5.4.1. **Descripción:** este programa permite estimar los parámetros de cuatro modelos de regresión ajustables a un par de variables

- | | | |
|-----|---------------------|------------------------|
| (1) | $Y = a + b x$ | lineal simple |
| (2) | $Y = a + e^{bx}$ | exponencial (semi-log) |
| (3) | $Y = a + b \log(x)$ | logarítmico |
| (4) | $Y = ax^b$ | potencial (log-log) |

Los modelos 2 y 4 que aparentemente son no lineales en los parámetros pueden linearizarse de la siguiente forma:

- | | | |
|-----|---------------------------------|----------|
| (2) | $\log(Y) = \log(a) + bX$ | |
| | $\log(Y) = a' + bX$ | semi-log |
| (4) | $\log(Y) = \log(a) + b \log(X)$ | |
| | $\log(Y) = a' + b \log(X)$ | log-log |
| | siempre y cuando $x \neq 0$ | |

4.5.4.2. **Funciones:** ONEVREG sirve para probar los cuatro modelos en un conjunto de datos con el fin de comparar los diferentes ajustes y, si es el caso, decidirse por uno de ellos como el modelo para representar el comportamiento que se está estudiando; desde luego cualquiera de ellos podría usarse con el programa MLINREG si se hacen previamente las transformaciones necesarias a las variables.

4.5.4.3. **Funcionamiento:** Los pasos generales en la utilización del programa son los siguientes:

- (1) Teclee ONEVREG <CR> en la primera instrucción STATPAK
- (2) El programa pregunta por el número de observaciones, indíquelo
- (3) Seleccione el (los) modelo(s) que le interesa probar respondiendo a STATPAK las preguntas pertinentes
- (4) Responda a las preguntas sobre archivo de entrada y asignación de campos
- (5) Indique el tipo de salida que desea

4.5.4.4. **Ejemplo:** Con los mismos datos usados en el ejemplo para el programa POLYREG (cuadro 4.63 y figura 4.64) se prueban los cuatro modelos de interés. El cuadro 4.67 muestra las instrucciones

requeridas y los resultados obtenidos con el programa. Con flechas numeradas se señalan los siguientes punto de interés:

- (1) El título del programa
- (2) Número de observaciones en el análisis
- (3) Se pide que haga los ajustes con los cuatro elementos
- (4) Definición del archivo de entradas
- (5) Se indica que produzca una salida impresa
- (6) Tabla de resultados. El programa presenta los valores de X, los esperados de Y bajo los cuatro modelos y los valores "Y observados"
- (7) Estimadores de los coeficientes A y B para los diferentes modelos
- (8) Errores estandar de los estimadores
- (8) Coeficiente de determinación (R^2) para los diferentes modelos

4.5.4.5. Una pequeña discusión sobre el ejemplo:

Los resultados obtenidos con los programas POLYREG y ONEVREG indican que los tres mejores modelos de ajuste a los datos son:

Lineal:	$Y = 1.74 + 1.07 X$,	$R^2 = 87.81\%$
Cuadrático:	$Y = -1.63 + 2.64 X - .1366 X^2$,	$R^2 = 97.77\%$
Logarítmico:	$Y = .02 + 4.99 \text{Log}_n(x)$,	$R^2 = 97.68\%$

Inicialmente puede dejarse de lado el ajuste lineal puesto que es claramente superado por los otros dos, lo cual era de esperarse puesto que la respuesta estudiada tiene una suave, pero clarísima, curvatura que el modelo lineal no es capaz de explicar; el problema es cuál de los otros dos modelos escoger. Desde luego se sabe que los datos fueron generados con un modelo logarítmico, pero si este hecho se desconoce es claro que cualquiera de los dos modelos sirve para explicar los datos siempre y cuando el rango de variación de X (espacio de exploración) no se altere. Si se consideran valores mayores de 10.0 en X solo un nuevo análisis podría indicar qué tan bien ajustaba, en ese nuevo espacio de exploración, la función polinomial cuadrática. Finalmente debe decirse que el conocimiento sobre los aspectos biológicos del fenómeno puede dar la respuesta definitiva: si se espera biológicamente que "Y" siga creciendo indefinidamente (aunque

Cuadro 4.67

----> ONE INDEPENDENT VARIABLE REGRESSION <----

MAXIMUM # of Data Pairs? 20

-- Select Desired Regression(s) --

Linear: $A+(B*X)$ (Y/N)? Y

Exponential: $A*EXP(B*X)$ (Y/N)? Y.

Logarithmic: $A+B*LOG(X)$ (Y/N)? Y

Power Curve: $A*X^B$ (Y/N)? Y

X = Field # 1

Y = Field # 2

Input File (Q=Quit, <CR>=ONEV.DAT):

2 Data Field(s)

2 File Columns in ONEV.DAT

Field Assignment(s) (<CR>=1:1):

.....

Output : S(creen), P(rinter), F(ile), or Q(uit)? P

X	A+(B*X)	A*EXP(B*X)	A+B*LOG(X)	A*XTB	Y
1	2.814983	1.966164	1.892233E-02	.7877131	.1400209
1.5	3.349551	2.219426	2.045001	1.340303	1.660309
2	3.884119	2.505311	3.482527	1.954262	3.635089
2.5	4.418686	2.828021	4.597559	2.618307	4.871137
3	4.953254	3.1923	5.508606	3.3252	5.373233
3.5	5.487822	3.603501	6.278886	4.069839	6.009323
4	6.022389	4.067669	6.946131	4.848392	6.514935
4.5	6.556957	4.591626	7.534684	5.657865	6.978947
5	7.091524	5.183076	8.061163	6.495841	8.456581
5.5	7.626092	5.850709	8.537422	7.360314	8.686319
6	8.16066	6.604342	8.97221	8.249594	9.880029
6.5	8.695227	7.455047	9.372177	9.162238	9.506448
7	9.229794	8.415335	9.742491	10.09699	10.16076
7.5	9.764362	9.499316	10.08724	11.05275	10.43481
8	10.29893	10.72293	10.40974	12.02853	11.32288
8.5	10.8335	12.10415	10.71267	13.02347	10.23935
9	11.36807	13.66329	10.99829	14.03678	9.700434
9.5	11.90263	15.42327	11.26846	15.06775	11.14074
10	12.4372	17.40994	11.52477	16.11574	11.25653
10.5	12.97177	19.65252	11.76857	17.18014	11.89964
A REG COEFF	1.745848	1.543042	1.892233E-02	.7877131	
B REG COEFF	1.069135	.2423286	4.996926	1.310882	
A STD ERROR	.6038091	.5675908	.3091507	.2432787	
B STD ERROR	9.387083E-02	5.718585E-02	.1813322	.181151	
STD ERR EST	1.210352	.7373426	.5276121	.5270849	
COEFF DET	.8781471	.4994011	.9768451	.7441931	
COVARIANCE	9.354932	2.120375	2.226532	.5841034	
CORR COEFF	.9370951	.7066832	.9883548	.8626663	
DURBN-WATSN	.4088845	.1246878	1.493177	.1166049	



a tasa decreciente) el modelo apropiado es el logarítmico; si se espera que "Y" alcance un máximo y luego permanezca estable o decrezca el modelo apropiado es el polinomial (de hecho $X_{\text{máx}} = 9.66$ para este caso).

La figura 4.68 presenta la gráfica de los valores estimados de "Y" para los dos modelos contra los valores de X (1 : logarítmico, 2 = cuadrático)

ID: LOGARITMICO VS CUADRATICO

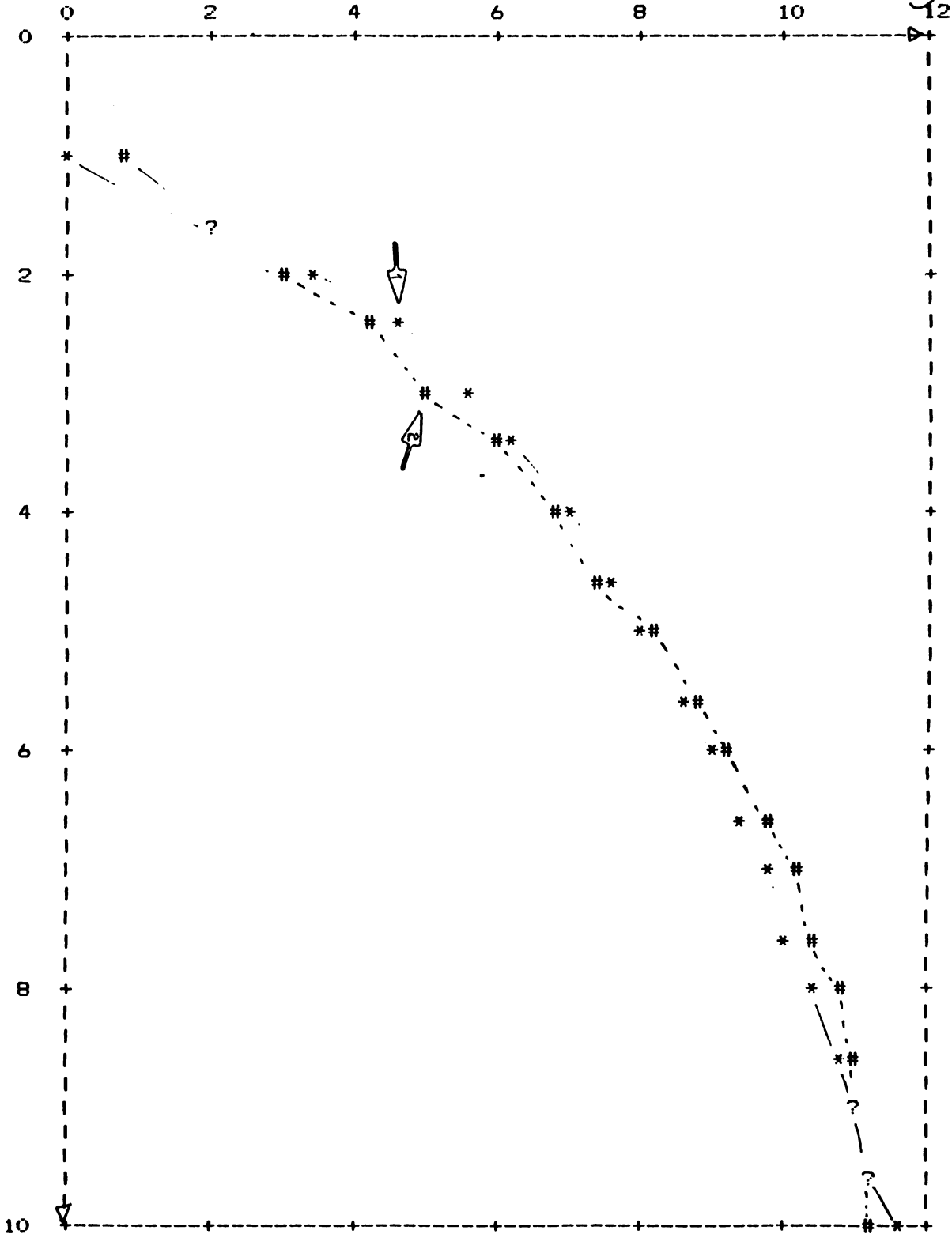
Date:

X * 1

Y * 1

COLLISION MARK: ?

FILE: WORK.DAT Columns: X = 1 (*) = 4 (#) = 12



4.6 Programas para Graficación

4.6.1 QPLOT

Elaboración rápida de gráficos (QUICK PLOT).

4.6.1.1. Descripción: QPLOT elabora gráficos de puntos, sujeto a las restricciones de una impresora corriente, eligiendo el mismo, automáticamente, todas las características de formato.

4.6.1.2. Funciones: Con QPLOT pueden elaborarse gráficos a partir de un único archivo dando opción a que se tengan hasta cuatro variables en el eje de las ordenadas y una, con valores en el archivo o generados por el programa, en el eje de las abscisas. Las gráficas de QPLOT presentan el eje de abscisas en sentido vertical y el de ordenas en el horizontal con el fin de permitir graficar en impresoras de papel ancho y también en aquellas que utilizan papel angosto. El programa exige que los valores de X estén ordenados en forma ascendente.

4.6.1.3. Funcionamiento: los pasos en el manejo de QPLOT son los siguientes:

- (1) digite QPLOT <CR > en la primera instrucción STATPAK
- (2) indique el número de valores "Y" (ordenada) que precisa graficar (todos contra el mismo conjunto de valores de X)
- (3) defina el archivo de entrada
- (4) la asignación de campos se hace variable por variable, iniciando con "X"
- (5) señale el tipo de salida que desea
- (6) si indica P (salida por impresora) STATPAK pregunta por la identificación de la salida

4.6.1.4. Ejemplo: en la figura 4.70 se presenta el gráfico de los valores "Y" utilizados como ejemplo para el procedimiento ONEVREG, cuyos datos aparecen en el cuadro 4.63. El cuadro 4.69 muestra las instrucciones utilizadas para elaborar la gráfica pidiendo a la máquina la generación de los valores X. Señalados con flechas numeradas aparecen los siguientes puntos:

- (1) STATPAK advierte que solo pueden graficarse cuatro conjuntos "Y" simultáneamente



Cuadro 4.69

----> QUICK PLOT <----

Maximum # of Y fields: 4 



Data Field 1 = X-values

Data Fields 2 - 5 = Y-values


How many Y fields to plot (<CR>=1,0=Quit)? 1 Input File (Q=Quit, <CR>=WORK.DAT): ONEV.DAT 

2 File Columns in ONEV.DAT

(Enter 0 to Auto-Generate X)

File Field for X values (<CR>= 1): 0 First X-Value (<CR>=0): 1 

Delta X (<CR>=10): .5-.6

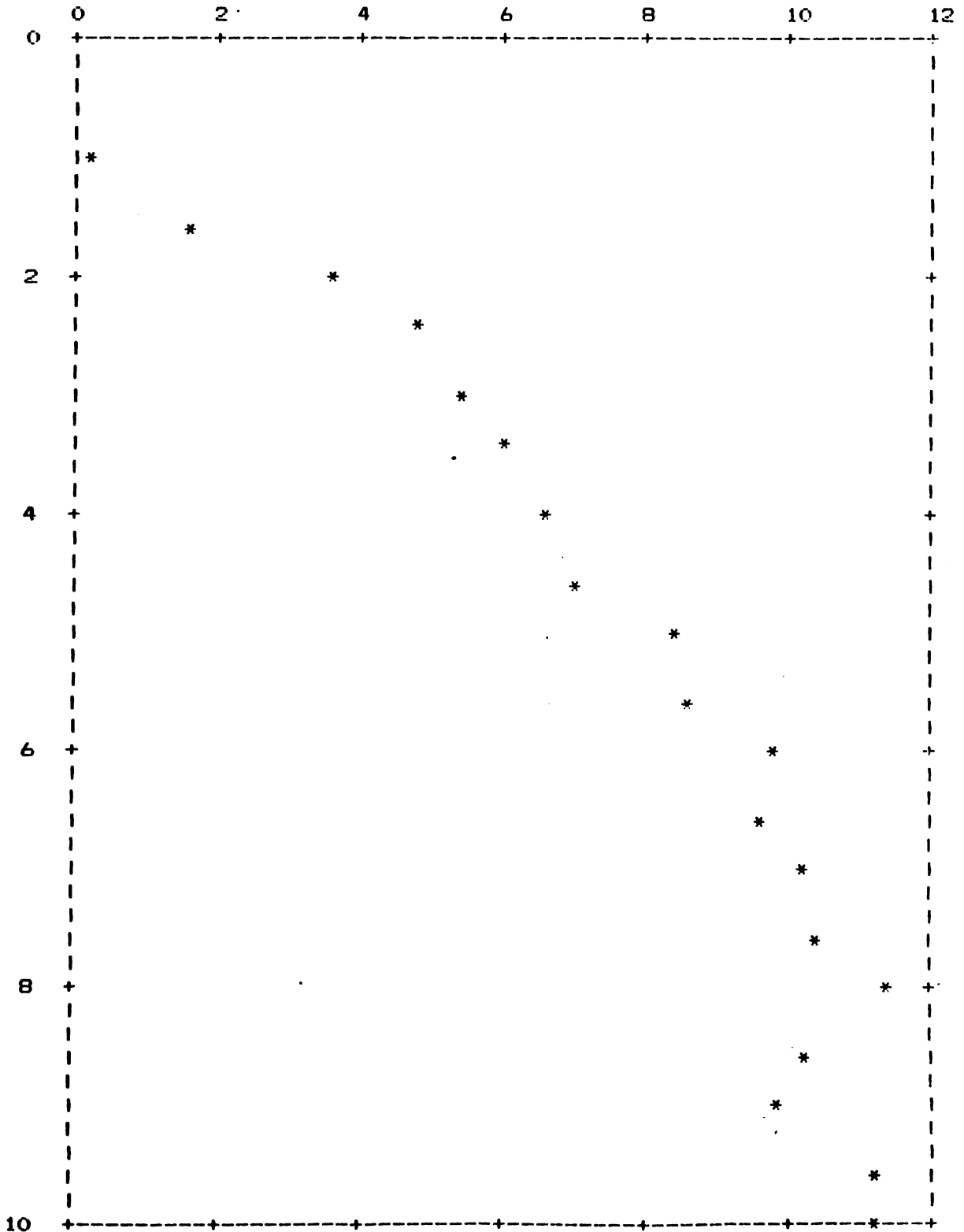
File Field for Y 1 values (<CR>= 1): 2 

Finding limits.....

Output : S(creen), P(rinter), F(ile), or Q(uit)? P 

X * 1 Y * 1 COLLISION MARK: ?

FILE: ONEV.DAT Columns: X = 0 (*) = 2



- (2) a la pregunta de cuántas variables "Y" se quiere graficar se responde que una
- (3) los valores a graficar están en el archivo ONEV.DAT
- (4) se pide a STATPAK que genere los valores para X...
- (5) ... tomando $X = 1$ como valor inicial...
- (6) ... y un incremento de .5 entre valores
- (7) se indica que los valores "Y" son la segunda columna de ONEV.DAT (puesto que son 20 valores QPLOT genera 20 valores para X y los asigna uno a uno)
- (8) finalmente se pide que salgan los resultados por la impresora

4.6.2. PLOT

Programa para graficación de datos y/o funciones

4.6.2.1. Descripción: PLOT elabora gráficos de puntos, sujetos a las restricciones de una impresora común, permitiendo al usuario determinar los formatos de salida y la graficación tanto de datos contenidos en archivos como de aquellos generados por el mismo programa de acuerdo a alguna función determinada por el usuario.

4.6.2.2. Funciones: PLOT permite graficar simultáneamente cinco conjuntos de datos y cinco funciones generadas (con valores esperados a partir de una ecuación estimada, por ejemplo) en la misma gráfica. Las funciones pueden ser de tipo lineal ($Y=A+Bx$), exponencial ($A+e^{Bx}$), logarítmica ($Y = A + B \text{Log}(x)$), potencia ($Y=Ax^B$) y polinomial ($Y = A + Bx^2 + Cx^3 + \dots$) con grado 3 como mínimo.

El usuario define los límites para el gráfico (valores mínimo y máximo de los dos ejes), la longitud de los ejes (si no quiere los que usa PLOT que corresponden a papel tamaño carta), los símbolos para representar los puntos dentro de la gráfica y el formato de los valores de identificación sobre los ejes. Poder definir las anteriores características implica que el usuario tenga claridad sobre los gráficos que está elaborando.

4.6.2.3 Ejemplo: con el fin de ejemplificar el uso de PLOT se toman los datos de ONEVREG (ONEV.DAT) y se presenta una gráfica con los valores observados de "Y" y los esperados bajo los dos mejores modelos: logarítmico y cuadrático, con base en las ecuaciones estimadas por el programa ONEVREG:

$$Y = .02 + 4.99 \text{Log}(x), \text{ y } Y = -1.63 + 2.64x - .137 x^2$$

El cuadro 4.71 presenta las instrucciones que se dieron a STATPAK y la figura 4.72, el gráfico resultante. Señalados, con flechas numeradas, aparecen los siguientes puntos de interés:

- (1) se quiere graficar un archivo...
- (2) ... y dos funciones
- (3) el nombre del archivo
- (4) en ONEV.DAT la primera columna contiene los valores "X", la segunda los "Y"
- (5) PLOT calcula los límites para "X" e "Y" en el archivo y los indica
- (6) se define como primera función la logarítmica con coeficientes .02 y 4.99 y se pide que los puntos sean marcados con asteriscos (*)
- (7) como segunda función se define una polinomial de grado 3 con coeficientes -1.63, 2.64, -.1366 y 0.0 para el término cúbico (la función de interés es de segundo orden y PLOT exige en el parámetro de orden un valor de 3 a 9, por esto se usa ese artificio)
- (8) se establecen límites para el gráfico:
abcisas: de 0.0 a 11.0
ordenadas: de - .50 a 12.0
- (9) el tamaño del gráfico se deja como PLOT lo ha definido (papel tamaño carta)
- (10) se definen formatos para los identificadores "X" e "Y" en el gráfico
- (11) se indica que si en un mismo punto se encuentran dos valores ese punto se marque con el símbolo
- (12) salida por impresora @

Cuadro 4.71

```

      ---> PLOT <---

Data Column #1: X-values
Data Column #2: Y-values

Maximum # of files: 5
Maximum # of functions: 5

How many files to plot? 1 ←1
How many functions? 2 ←2

--> File # 1
Plot Symbol: .
Input File (Q=Quit, <CR>=): ONEV.DAT ←3
  2 File Columns in ONEV.DAT
  (Enter 0 to Auto-Generate X)
File Column for X values: 1 ←4
File Column for Y values: 2

File Column for Y values: 2
Finding limits.....
--- FILE 1 ---
Min X: 1           Min Y: .1400209 ←5
Max X: 10.5       Max Y: 11.89964

--> Function # 1
Plot Symbol: * ←6
1=Linear 2=Exponential 3=Logarithmic 4=Power 5=Polynomial
A+B*X    A*e↑(B*X)    A+B*LOG(X)    A*X↑B    B0+B1*X+B2*X↑2...

Which type (1-5)? 3 ←6
A Value? .02
B Value? 4.99

--> Function # 2
Plot Symbol: + ←3
1=Linear 2=Exponential 3=Logarithmic 4=Power 5=Polynomial
A+B*X    A*e↑(B*X)    A+B*LOG(X)    A*X↑B    B0+B1*X+B2*X↑2...

Which type (1-5)? 5

```

Cuadro 4.71 (continuación)

A+B*X A*e↑(B*X) A+B*LOG(X) A*X↑B B0+B1*X+B2*X↑2...

Which type (1-5)? 5

Order of Polynomial (3-9)? 3

Coefficient 0 ? -1.63

Coefficient 1 ? 2.64

Coefficient 2 ? -.1366

Coefficient 3 ? 0



---> ESTABLISH PLOT LIMITS <---

X Data min = 1

Plot min (<CR>= 1): 0

X Data max = 10.5

Plot max (<CR>= 10.5): 11

Y Data min = .1400209

Plot min (<CR>= .1400209): -.5

Y Data max = 11.89964

Plot max (<CR>= 11.89964): 12

X-axis length (min=10; <CR>= 50): 9

Y-axis length (min=10; <CR>= 60): 8

Print Extra Tick Marks (Y/N)? N

X-Axis Tick Label Picture (<CR>=default): ##.##

Y-Axis Tick Label Picture (<CR>=default): ##.#####

Collision Plot Symbol: @

Output: S(creen), P(rinter), F(ile), or Q(uit)? P

DX = .22

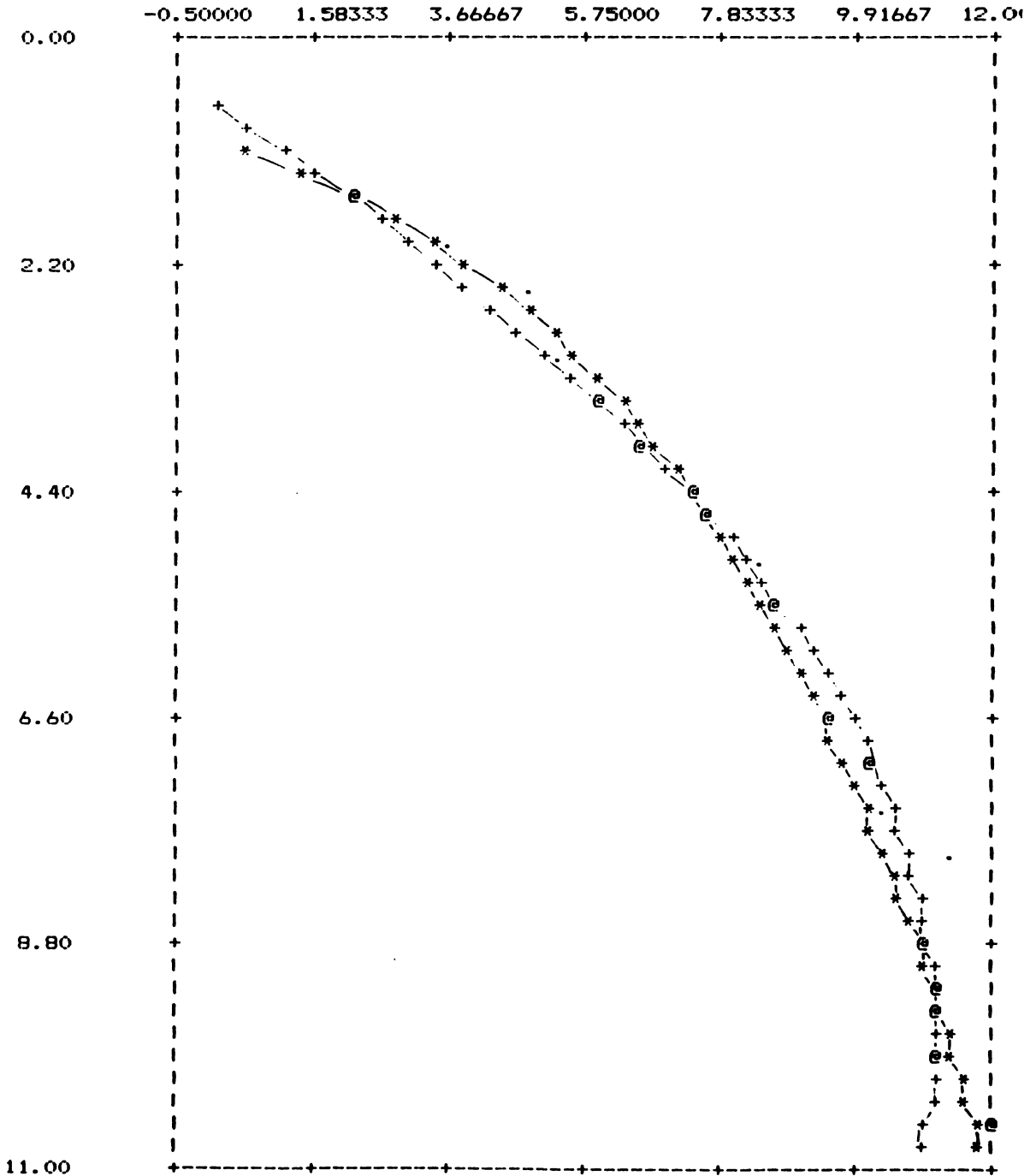
DY = .2083333

COLLISION SYMBOL: e

FILES: ONEV.DAT (1,2)=.

FUNCTIONS: * = .02 + 4.99*LOG(X)

+ = -1.63 + 2.64X - .1366X^2 + 0X^3



4.7 Programas para el análisis de diseños experimentales

4.7.1. Descripción: STAPAK presenta ocho programas para el análisis de datos obtenidos a partir de algún diseño experimental. En opinión de quien escribe esta parte del STATPAK es la menos elaborada y la que presenta algunos problemas al usuario. No obstante una utilización inteligente de los diferentes programas permitirá analizar la mayoría de los diseños comunes y la combinación de experimentos en el tiempo y/o en el espacio.

Una primera carencia en estos programas es que no calculan promedios, pero estos pueden tenerse haciendo algunos cambios sencillos en los programas (de hecho dos de ellos ya han sido modificados por quien escribe para que presenten las medias de tratamientos: ANOVA1R y ANOVA2R2).

4.7.2. Programas: los programas se listan en seguida junto con los diseños que pueden analizarse con ellos, usando directamente sus resultados o calculando los items requeridos a partir de los resultados que el programa ofrece:

- ANOVA1: diseño completamente al azar, DCA
- ANOVA1R: bloques completos al azar, DBCA
- ANOVA2: bloques al azar con muestreo en las unidades experimentales. Factorial A x B en DCA
- ANOVA2R1: parcelas divididas en DCA. Factorial A x B en DCA*
- ANOVA2R2: factorial A x B en DBCA*. Parcelas divididas en DBCA*. Combinado de experimentos en DBCA con igual número de repeticiones*
- PDIV: es una modificación al anterior programa hecha por quien escribe, para tener la salida clásica de parcelas divididas en DBCA, incluyendo medias
- ANOVA3: factorial A x B x C en DCA. Parcelas divididas en DBCA con muestreo en las unidades experimentales*
- ANOVA3R1: parcelas divididas con parcelas principales formadas por la combinación de niveles de dos factores A, B en DCA
- ANOVA3R2: factorial A x B x C en DCA*. Factorial B x C en DBCA con muestreo en las unidades experimentales*

*. Se requieren cálculos adicionales a partir de los resultados del programa

4.7.3. Estructura de los datos

El segundo problema que presentan los programas para diseño experimental de STATPAK está originado en el hecho de que cada programa exige un tipo de archivo con estructura diferente de los datos. Este problema hace difícil el análisis de los mismos datos usando diferentes programas, caso de parcelas divididas en el tiempo o del análisis conjunto de experimentos. Las diferentes estructuras de datos se definen adelante. Es importante notar que las variables de clasificación o identificación no son necesarias en los archivos de datos, basta con tener la estructura adecuada para que los programas puedan operar.

4.7.3.1. ANOVA 1: (DCA)

Hileras: Repeticiones

Columnas: Tratamientos

$$\text{Modelo: } Y_{ij} = \mu + \tau + \epsilon_{ij}$$

$$j = 1, 2, \dots, t$$

$$i = 1, 2, \dots, r_j$$

T_1	T_2	...	T_t
Y_{11}	Y_{12}	...	Y_{1t}
Y_{21}	Y_{22}	...	Y_{2t}
.	.	.	.
.	.	.	.
Y_{r_1}	Y_{r_2}		Y_{r_t}

con r_1, r_2, \dots, r_t iguales o diferentes. En caso de que sean diferentes las columnas deben pertenecer a diferentes archivos puesto que STATPAK no admite archivos con datos faltantes.

4.7.3.2. ANOVA IR: (DBCA)

Hileras: Bloques

Columnas: Tratamientos

$$\text{Modelo: } Y_{ij} = \mu + \rho_i + \tau_j + \epsilon_{ij}$$

$$i = 1, 2, \dots, r$$

$$j = 1, 2, \dots, t$$

TRATAMIENTOS

	T_1	T_2	...	T_t	
B L O Q U E S	B_1	Y_{11}	Y_{12}	...	Y_{1t}
	B_2	Y_{21}	Y_{22}	...	Y_{2t}

	B_r	Y_{r1}	Y_{r2}	...	Y_{rt}

4.7.2.3. ANOVA2: Factorial A x B en DCA

Hileras: cada observación forma una hilera

Columnas: un sola columna por variable

Modelo: $Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \epsilon_{ijk}$

$i = 1, 2, \dots, a$ (niveles de A)

$j = 1, 2, \dots, b$ (niveles de B)

$k = 1, 2, \dots, r$ (repeticiones)

En el caso de un factorial 3 x 2 con dos repeticiones el archivo tendrá la forma:

	B_1	Y_{111} Y_{112}
A_1		
	B_2	Y_{121} Y_{122}
	B_1	Y_{211} Y_{212}
A_2		
	B_2	Y_{221} Y_{222}
	B_1	Y_{311} Y_{312}
A_3		
	B_2	Y_{321} Y_{322}

Esto es tendrá 12 hileras y una (si solo una variable) o más columnas según el número de variables que se quieran analizar.

Note que si en el modelo descrito se toma A como bloques, B como tratamientos y se supone que en cada parcela se tomaron muestras el modelo corresponde a bloques al azar con submuestreo, en donde $A \times B$ es el error experimental y "error residual" es el error de muestreo.

4.7.2.4. ANOVA2R1: Parcelas divididas en DCA

Hileras: repeticiones ordenadas por parcela principal

Columnas: subparcelas

Modelo: $Y_{ijk} = \mu + A_i + \delta_{ik} + B_j + AB_{ij} + \epsilon_{ijk}$

$i = 1, 2, \dots, a$ (parcelas principales)

$j = 1, 2, \dots, b$ (subparcelas)

$k = 1, 2, \dots, r$ (repeticiones)

Sea un diseño con 3 parcelas principales, 2 subparcelas y 3 repeticiones; el archivo de datos tendrá la forma:

		SUBPARCELA	
		B ₁	B ₂
A ₁		Y ₁₁₁	Y ₁₂₁
		Y ₁₁₂	Y ₁₂₂
		Y ₁₁₃	Y ₁₂₃

PARCELA PRINCIPAL A ₂		Y ₂₁₁	Y ₂₂₁
		Y ₂₁₂	Y ₂₂₂
		Y ₂₁₃	Y ₂₂₃

A ₃		Y ₃₁₁	Y ₃₁₁
		Y ₃₁₂	Y ₃₁₂
		Y ₃₁₃	Y ₃₁₃

Observe que si se suman los dos errores δ y ϵ en el modelo (desde luego esto es permitido si la aleatorización no fue en parcelas divididas, de otra forma no) para formar un solo error, se tiene un factorial A x B en DCA.

4.7.2.5. ANOVA2R2: Factorial A x B en DBCA. El programa PDIV tiene la misma estructura de datos puesto que se generó a partir de ANOVA2R2.

Hileras: Bloques

Columnas: Combinaciones de niveles A x B (una columna por combinación)

$$\text{Modelo: } Y_{ijk} = \mu + \rho_i + A_j + B_k + AB_{jk} + \epsilon_{ijk}$$

$$i = 1, 2, \dots, r \quad (\text{bloques})$$

$$j = 1, 2, \dots, a \quad (\text{niveles de A})$$

$$k = 1, 2, \dots, b \quad (\text{niveles de B})$$

Sea un factorial 3 x 2 con 3 repeticiones;
la estructura del archivo de datos será:

	A ₁		A ₂		A ₃	
	B ₁	B ₂	B ₁	B ₂	B ₁	B ₂
R ₁	Y ₁₁₁	Y ₁₁₂	Y ₁₂₁	Y ₁₂₂	Y ₁₃₁	Y ₁₃₂
R ₂	Y ₂₁₁	Y ₂₁₂	Y ₂₂₁	Y ₂₂₂	Y ₂₃₁	Y ₂₃₂
R ₃	Y ₃₁₁	Y ₃₁₂	Y ₃₂₁	Y ₃₂₂	Y ₃₃₁	Y ₃₃₂

En el programa PDIV A corresponde a las parcelas principales y B a las subparcelas.

4.7.2.6. ANOVA3: Factorial AxBxC en DCA

Hileras: Combinación de niveles de A con B y repeticiones

Columnas: Subparcelas

$$\text{Modelo: } Y_{ijkl} = \mu + A_i + B_j + AB_{ij} + C_k + AC_{ik} + BC_{jk} + ABC_{ijk} + \epsilon_{ijkl}$$

La estructura es similar a la de ANOVA2R1, pero cada parcela principal está ahora formada por la combinación de niveles de dos factores A y B.

4.7.3. Ejemplos: Los ejemplos a presentar utilizan solo tres de los programas disponibles: ANOVA1R, ANOVA2R2 y PDIV, que parecen ser los más utilizados. Se pide al lector que precise de otros ejemplos buscarlos en el manual original de STATPAK.

El problema ejemplo corresponde a una pequeña parte de los datos de una investigación realizada por Fernando López, estudiante de agronomía, en la Estación Experimental del Norte, CIAAB - Tacuarembó. El cuadro 4.73 presenta los contenidos de humedad en el suelo para dos profundidades diferentes (0-20 cm y 20-40 cm), en dos períodos de crecimiento vegetativo del cultivo y para cinco diferentes densidades de siembra en soja (T₁ la menor, T₅ la mayor). El diseño experimental básico es bloques al azar con cuatro bloques. Los títulos que aparecen con los datos fueron puestos para clarificar el ejemplo pero "NO FORMAN PARTE DEL ARCHIVO DE DATOS".

4.7.3.1. El primer paso es analizar los cuatro subgrupos de datos con el diseño de bloques al azar, programa ANOVA1R.

Cuadro 4.73

	PERIODO 1					PERIODO 2					
	T1	T2	T3	T4	T5	T1	T2	T3	T4	T5	
I	49.13	58.86	49.34	48.58	49.73	42.36	44.29	42.54	38.44	42.08	PROFUNDIDAD 0-20
II	48.87	49.55	46.69	46.89	45.51	30.00	36.75	39.50	39.02	36.11	
III	48.90	48.17	44.69	45.49	43.34	49.04	44.86	43.41	44.89	44.20	
IV	47.58	44.16	48.81	45.66	42.26	47.96	39.03	45.23	41.84	39.19	
I	53.87	50.23	49.74	49.45	50.99	42.75	43.60	42.48	41.28	46.73	PROFUNDIDAD 20-40
II	49.20	49.99	50.28	47.84	45.82	33.81	40.66	42.77	42.24	45.70	
III	51.31	49.66	50.50	48.38	53.55	43.19	44.40	45.14	46.81	49.21	
IV	52.71	49.10	53.50	52.25	46.18	52.00	49.28	47.09	48.05	45.93	

Cuadro 4.74

ANOVA1. DAT: PROFUNDIDAD 0-20										
I	49.13	58.86	49.34	48.58	49.73	42.36	44.29	42.54	38.44	42.08
II	48.87	49.55	46.69	46.89	45.51	30.00	36.75	39.50	39.02	36.11
III	48.90	48.17	44.69	45.49	43.34	49.04	44.86	43.41	44.89	44.20
IV	47.58	44.16	48.81	45.66	42.26	47.96	39.03	45.23	41.84	39.19

ANOVA2. DAT: PROFUNDIDAD 20-40										
I	53.87	50.23	49.74	49.45	50.99	42.75	43.60	42.48	41.28	46.73
II	49.20	49.99	50.28	47.84	45.82	33.81	40.66	42.77	42.24	45.70
III	51.31	49.66	50.50	48.38	53.55	43.19	44.40	45.14	46.81	49.21
IV	52.71	49.10	53.50	52.25	46.18	52.00	49.28	47.09	48.05	45.93

Cuadro 4.75 a

---> 1-WAY ANOVA: RANDOMIZED/ONE TREATMENT <---

Number of Treatments? 5

Input File (Q=Quit, <CR>=): ANOVA1.DAT

5 Data Field(s)

11 File Columns in ANOVA1.DAT

Field Assignment(s) (<CR>=1:1): 2 3 4 5 6

....

Output: S(screen), P(printer), F(file), or Q(quit)? P

--> MAKE PRINTER READY

Report ID: EJEMPLO1: BLOQUES AL AZAR, PROFUNDIDAD 0-20, PERIODO 1

---> PROG:ANOVA1R FILE:ANOVA1.DAT ID:EJEMPLO1: BLOQUES AL AZAR, PROFUNDIDAD 0-20, PERIODO 1

	SUM SQUARES	DEG FREEDOM	MEAN SQUARE
TREATMENT	57.50391	4	14.37598
BLOQUES	91.42578	3	30.47526
ERROR	78.5	12	6.541667
TOTAL	227.4297		

F-TEST RATIO: 2.197602

No TRAT	PROMEDIO
1	48.62
2	50.185
3	47.3825
4	46.65501
5	45.21

Cuadro 4.75 b

----> 1-WAY ANOVA: RANDOMIZED/ONE TREATMENT <----

Number of Treatments? 5

Input File (Q=Quit, <CR>=ANOVA1.DAT):

5 Data Field(s)

11 File Columns in ANOVA1.DAT

Field Assignment(s) (<CR>=:1): 6 7 8 9 10 11-4

....
Output: S(screen), P(rinter), F(ile), or Q(uit)? P

--> MAKE PRINTER READY

Report ID: EJEMPLO1: BLOQUES AL AZAR, PROFUNDIDAD 0-20, PERIODO 2

---> PROG:ANOVA1R FILE:ANOVA1.DAT ID:EJEMPLO1: BLOQUES AL AZAR, PROFUNDIDAD 0-20, PERIODO 2 PAGE 1

	SUM SQUARES	DEG FREEDOM	MEAN SQUARE
TREATMENT	44.45703	4	11.11426
BLOQUES	136.0078	3	45.33594
ERROR	216.3828	12	18.0319
TOTAL	396.8477		

F-TEST RATIO: .6163664

No TRAT	PROMEDIO
1	45.21
2	42.34
3	41.2325
4	42.67
5	41.0475

----> 1-WAY ANOVA: RANDOMIZED/ONE TREATMENT <----

Number of Treatments? 5

Input File (Q=Quit, <CR>=): ANOVA2.DAT

5 Data Field(s)

11 File Columns in ANOVA2.DAT

Field Assignment(s) (<CR>=1:1): 2 3 4 5 6

....

Output: S(screen), P(printer), F(file), or Q(uit)? P

--> MAKE PRINTER READY

Report ID: EJEMPLO1: BLOQUES AL AZAR, PROFUNDIDAD 20-40, PERIODO 1

--> PROG:ANOVA1R FILE:ANOVA2.DAT ID:EJEMPLO1: BLOQUES AL AZAR, PROFUNDIDAD 20-40, PERIODO 1 PAGE 1

	SUM SQUARES	DEG FREEDOM	MEAN SQUARE
TREATMENT	19.90625	4	4.976562
BLOQUES	17.17188	3	5.723959
ERROR	58.48828	12	4.874024
TOTAL	95.56641		

F-TEST RATIO: 1.021038

No TRAT	PROMEDIO
1	51.7725
2	49.74501
3	51.005
4	49.48
5	49.135

Cuadro 4.75 d

---> 1-WAY ANOVA: RANDOMIZED/ONE TREATMENT <---

Number of Treatments? 5
Input File (Q=Quit, <CR>=ANOVA2.DAT):
5 Data Field(s)
11 File Columns in ANOVA2.DAT
Field Assignment(s) (<CR>=1:1): 7 8 9 10 11
....
Output: S(screen), P(printer), F(file), or Q(quit)? P
--> MAKE PRINTER READY
Report ID: EJEMPLO1:BLOQUES AL AZAR, PROFUNDIDAD 20-40, PERIODO 2

--> PROG:ANOVA1R FILE:ANOVA2.DAT ID:EJEMPLO1:BLOQUES AL AZAR, PROFUNDIDAD 20-40, PERIODO 2
PAGE 1

	SUM SQUARES	DEG FREEDOM	MEAN SQUARE
TREATMENT	32.28125	4	8.070313
BLOQUES	152.5352	3	50.84505
ERROR	106.6641	12	8.888672
TOTAL	291.4805		

F-TEST RATIO: .9079323

NO TRAT	PROMEDIO
1	42.9275
2	44.485
3	44.37
4	44.595
5	46.8925

Para ello se separaron los datos en dos archivos: ANOVA1.DAT y ANOVA2.DAT (cuadro 4.74) para cumplir con los requisitos de estructura exigidos por el programa. Los cuadros 4.75(a) a 4.75(d) muestran las instrucciones dadas a STATPAK y los resultados obtenidos. Con flechas numeradas se señalan los siguientes puntos:

- (1) Luego de dar la instrucción ANOVAIR CR en la primera instrucción STATPAK el programa pregunta por el número de tratamientos (columnas) a analizar, no pregunta por el número de bloques puesto que este corresponde al número de hileras
- (2) En los dos primeros casos se pide analizar ANOVA1.DAT, en los siguientes ANOVA2.DAT
- (3) El programa avisa que se requieren 5 columnas para cada análisis y que los archivos tienen 11 columnas cada uno (una correspondiente al número de bloque, la primera, y 10 a los 5 tratamientos en los dos períodos), observe los datos en el cuadro 4.74
- (4) Se le indica cuáles columnas entran en cada oportunidad al análisis. Observe detenidamente este punto pues indica cómo analizar parcialmente los dos conjuntos de datos
- (5) Se pide salida impresora
- (6) Definición del título para la salida
- (7) Dentro de los resultados es interesante, para el ejemplo notar los cuadrados medios del error

Prof. 0-20, Período 1 =	6.54
Prof. 0-20, Período 2 =	18.03
Prof. 20-40, Período 1 =	4.87
Prof. 20-40, Período 2 =	8.89

4.7.3.2. Como segundo paso en el análisis se plantea la comparación entre las dos profundidades considerando cada período por separado. Si se toma en cuenta que el suelo a 0-20 cm es diferente de aquel a 20-40 cm se puede conceptualmente pensar que en cada profundidad hay un experimento, con los mismos tratamientos, pero con diferentes bloques físicos y entonces la comparación entre profundidades debe hacerse mediante un análisis conjunto con el modelo:

$$Y_{ijk} = \mu + \rho_k + A_i + B_j + AB_{ij} + \epsilon_{ijk}$$

$i = 1, 2$ (profundidades)
 $j = 1, 2, 3, 4, 5$ (tratamientos)
 $k = 1, 2, 3, 4$ (bloques en cada profundidad)

El cual conduce a una tabla de análisis de varianza como la siguiente:

F de	GL
Bloques (profundidad)	$a(r - 1) = 2(3) = 6$
Profundidad (A)	$(a - 1) = 2 - 1 = 1$
Tratamientos (B)	$(b - 1) = 5 - 1 = 4$
Prof x Trat (AxB)	$(a - 1)(b - 1) = 4$
Error	$a(r - 1)(b - 1) = 2(3)(4) = 24$
Total	$abr - 1 = 39$

Permitido siempre y cuando las varianzas de los errores en los dos experimentos a combinar sean "estadísticamente iguales". Puesto que en el ejemplo se trata de dos experimentos una prueba de F es suficiente para probar la igualdad de varianzas:

en el período 1: (ver cuadros 4.75(a) a 4.75(d), N° 7)

$$F = \frac{6.5417}{4.8740} = 1.34 \text{ N.S}$$

en el período 2:

$$F = \frac{18.0319}{8.889} = 2.03 \text{ N.S}$$

$F(.05, 12, 12) = 2.69$, así, es posible hacer los análisis combinados.

Para el ejemplo se considera sólo el análisis combinado en el período 1. Con el fin de cumplir los requisitos del programa ANOVA2R2, que se utilizará para el análisis, se crea, usando el programa SPLICE, el archivo ANOVA3.DAT, a partir del archivo completo (cuadro 4.74). El cuadro 4.76 presenta los datos correspondientes, observe cuidadosamente su estructura y compárelo con el archivo inicial para una mejor comprensión.

El cuadro 4.77 muestra las instrucciones utilizadas. Con flechas numeradas se marcan los siguientes puntos:

Cuadro 4.76

	PROFUNDIDAD 0-20					PROFUNDIDAD 20-40				
I	49.13	58.86	49.34	48.58	49.73	53.87	50.23	49.74	49.45	50.99
II	48.87	49.55	46.69	46.89	45.51	49.20	49.99	50.28	47.84	45.82
III	48.90	48.17	44.69	45.49	43.34	51.31	49.66	50.50	48.38	53.55
IV	47.58	44.16	48.81	45.66	42.26	52.71	49.10	53.50	52.25	46.18

Cuadro 4.77

```

----> 2-WAY ANOVA: RANDOMIZED/TWO TREATMENT <----
.
Levels of Factor A ? 2 <1--
Levels of Factor B ? 5 <5--
Total Data Columns = 10 <3--


-- Data Columns: Factor B varies most rapidly --

Input File (Q=Quit, <CR>=WORK.DAT): ANOVA3.DAT <4--
  10 Data Field(s)
  11 File Columns in ANOVA3.DAT <5--
Field Assignment(s) (<CR>=1:1): 2 3 4 5 6 7 8 9 10 11 <6--

....
  4 Subjects read.

Output : S(screen), F(rinter), F(ile), or Q(uit)? F
--> MAKE PRINTER READY
Report ID: EJEMPLO 2: ANALISIS CONJUNTO, PERIODO1. A=PROFUNDIDAD,
                                             B=TRATAMIENTO

```



- (1) El factor A tiene dos niveles (note que en el título, N° 7, se anota A: Profundidad)
- (2) El factor B tiene cinco niveles (B: Tratamiento)
- (3) El programa avisa que requiere 10 columnas para el análisis
- (4) Se define el archivo a analizar
- (5) ANOVA2R2 anuncia que ANOVA3.DAT tiene 11 columnas y él requiere solo 10
- (6) Se le indica cuáles columnas (y en qué orden) tomar. Se elimina la columna 1 que tiene el número de bloque
- (7) Se pide salida impresa e indica el título

El cuadro 4.78 presenta los resultados del programa, con flechas numeradas señala lo siguiente:

- (1) Identificación del programa utilizado
- (2) Primer error: corresponde a la interacción bloque x factor A
- (3) Segundo error: interacción bloque x factor B
- (4) Tercer error: interacción bloque x factor A x factor B
- (5) Promedios para los factores y la interacción

A partir de estos resultados se construye la tabla del análisis de varianza conjunto de la siguiente forma:

- (1) Bloques (factor A): bloques + bloques x factor A
 $GL = 3 + 3 = 6$
 $SC = 57.83594 + 50.76563 = 108.60157$
- (2) Error = (Bloques x factor B) + (Bloques x factor A x factor B)
 $GL = 12 + 12 = 24$
 $SC = 81.28125 + 55.71094 = 136.99219$

Quedando la siguiente tabla:

--> PROG: ANOVA2R2 FILE: ANOVA3.DAT DATE: 06-30-1986 ID: EJEMPLO2: ANALISIS CONJUNTO, PERIODO 1.

A=PROFUNDIDAD, B=TRATAMIENTO

F-TEST RATIO

MEAN SQUARE

DEG FREEDOM

SUM SQUARES

	SUM SQUARES	DEG FREEDOM	MEAN SQUARE	F-TEST RATIO
BLOQUES	57.83594	3	19.27865	1.139274
FACTOR A	68.46094	1	68.46094	4.045706
ERROR	50.76563	3	16.92188	
FACTOR B	52.57813	4	13.14453	1.9406
ERROR	81.28125	12	6.773438	
A x B	24.82813	4	6.207032	1.336979
ERROR	55.71094	12	4.642578	

PROMEDIOS, FACTOR A

1	47.6105
2	50.2275

PROMEDIOS, FACTOR B

1	50.19625
2	49.965
3	49.19375
4	48.0675
5	47.1725

PROMEDIOS, INTERACCION

1	48.62
1	50.185
1	47.3825
1	46.65501
1	45.21
2	51.7725
2	49.74501
2	51.005
2	49.48
2	49.135



Cuadro 4.79

ANOVA4.DAT: PROFUNDIDAD 20-40

	T 1		T 2		T 3		T 4		T 5	
	PER	PER 2	PER	PER 2	PER	PER 2	PER	PER 2	PER	PER 2
I	53.87	42.75	50.23	43.60	49.74	42.48	49.45	41.28	50.99	46.73
II	49.20	33.81	49.99	40.66	50.28	42.77	47.84	42.24	45.82	45.70
III	51.31	43.19	49.66	44.40	50.50	45.14	48.38	46.81	53.55	49.21
IV	52.71	52.00	49.10	49.28	53.50	47.09	52.25	48.05	46.18	45.93

F de	GL	SC	CM	Fc
Bloques (Prof)	6	108.60157	18.10026	
Profundidades	1	68.46094	68.46094	11.99
Tratamientos	4	52.57813	13.14453	2.30
Prof x Trat	4	24.82813	6.20703	1.09
Error	24	136.99219	5.70800	
TOTAL	39	391.46096		

4.7.3.3. Finalmente se considera la comparación entre períodos de tiempo para cada una de las profundidades. Es claro que en este caso se tienen dos mediciones en el tiempo (período 1, período 2) en la misma unidad experimental y el modelo a considerar es el de parcelas divididas en el tiempo, tomando tratamientos como parcela principal y período como subparcela. El modelo estadístico es entonces:

$$Y_{ijk}: \mu + \rho_i + A_j + \theta_{ij} + B_k + AB_{jk} + \epsilon_{ijk}$$

$i = 1, 2, 3, 4$ (bloques)

$j = 1, 2, 3, 4, 5$, (tratamientos = parcela principal)

$k = 1, 2$ (períodos = subparcela)

Para ejemplificar se hará el análisis en la profundidad 20-40 cm. Los datos necesarios, ordenados para que sea claro el proceso se presentan en el cuadro 4.79. Es interesante compararlo con 4.74 puesto que es el mismo conjunto de datos, pero ordenado de otra forma. El cuadro 4.80 muestra las instrucciones indicadas, con flecha se señala:

- (1) Luego de dar la instrucción PDIV < CR > en la primera instrucción STATPAK el programa pregunta por el número de niveles del factor A, se responde que 5 (tratamientos)
- (2) El factor B (períodos) tiene dos niveles
- (3) El programa indica que requiere 10 columnas (5 x 2) para hacer el análisis
- (4) Se define ANOVA4.DAT como archivo de entrada
- (5) El programa indica que ANOVA4.DAT tiene 11 columnas y él sólo usará 10

(6) Se definen las columnas a usar

(7) Los resultados deberán salir impresos y con el título anotado

El cuadro 4.81, presenta finalmente, el análisis de las parcelas divididas y los promedios requeridos.

PROG:PDIV FILE:ANOVA4.DAT ID:EJEMPLO3: PARCELAS DIVIDIDAS EN EL TIEMPO, PROFUNDIDAD 20-40
PAGE 1

	SUM SQUARES	DEG FREEDOM	MEAN SQUARE	F-TEST RATIO
BLOQUES	122.1797	3	40.72656	4.441636
FACTOR A	5.3125	4	1.328125	.1448452
ERROR A	110.0313	12	9.169271	
FACTOR B	310.4063	1	310.4063	45.35962
A x B	46.875	4	11.71875	1.712459
ERROR B	102.6484	15	6.84323	
TOTAL	697.4531	39		

PROMEDIO = 47.44176

PROMEDIOS, FACTOR A	
1	47.355
2	47.115
3	47.6875
4	47.0375
5	48.01375
PROMEDIOS, FACTOR B	
1	50.2275
2	44.556
PROMEDIOS, INTERACCION	
1	51.7725
1	42.9275
2	49.74501
2	44.485
3	51.005
3	44.37
4	49.48
4	44.595
5	49.135
5	46.8925

**Esta publicación tiene un tiraje de 250 ejemplares
y se terminó de imprimir en la ciudad de Montevideo,
Uruguay, en el mes de diciembre de 1986.**

**Para su impresión y distribución se contó con
el aporte de la Dirección del Programa II Generación
y Transferencia de Tecnología.**

Autor: Jorge Franco Durán

**Levantamiento de textos y asistencia editorial:
Graciela Nario, Ana Picabea, Rosa Waldman,
Nicole Hornblas**

Impresión: La Mimeográfica, Paysandú 1815 bis

COMISION DEL PAPEL. Edición amparada en el Art. 79 de la Ley 13.349

El Programa de Generación y Transferencia de Tecnología es la respuesta del IICA a dos aspectos fundamentales: (i) el reconocimiento por parte de los países y de la comunidad técnico-financiera internacional de la importancia de la tecnología para el desarrollo productivo del sector agropecuario; (ii) el convencimiento generalizado de que para aprovechar plenamente el potencial de la ciencia y la tecnología es necesario que existan infraestructuras institucionales capaces de desarrollar las respuestas tecnológicas apropiadas a las condiciones específicas de cada país y un marco de políticas que promueva y posibilite que las mismas sean incorporadas a los procesos productivos.

En ese contexto, el Programa de Generación y Transferencia de Tecnología se orienta a promover y apoyar las acciones de los países miembros tendientes a mejorar el diseño de sus políticas tecnológicas, fortalecer la organización y administración de sus sistemas de generación y transferencia de tecnología y facilitar la transferencia internacional de tecnología. De tal modo se podrá lograr un mejor aprovechamiento de todos los recursos disponibles y una mejor y más efectiva contribución a la solución de los problemas tecnológicos de la producción agropecuaria, dentro de un marco de igualdad en la distribución de los beneficios y conservación de los recursos naturales.

PLAN DE MEDIANO PLAZO 1987 - 1991



FECHA DE DEVOLUCION

IICA
PM-A4/UY-86-004

Autor

Título Cómo usar el paquete estadístico STATPAK

Nombre del solicitante

Fecha Devolución	



