Estadística aplicada al Turismo utilizando SPSS y STATGRAPHICS Plus

Ángela María González Laucirica, MSc.

Universidad Ecotec
Formación para el buen vivir

Guayaquil - Ecuador

Universidad Ecotec
Formación para el buen vivir

TÍTULO
Estadística aplicada al Turismo utilizando SPSS y STATGRAPHICS Plus.

AUTORA
Ángela María González Laucirica, MSc.

AÑO
2013

EDICIÓN
Centro de Publicaciones - Universidad ECOTEC

ISBN
978-9978-9931-8-7

No. DE PÁGINAS
256

LUGAR DE EDICIÓN
Guayaquil - Ecuador

DISEÑO DE CARÁTULA
DAGMAR

TIRAJE
500 ejemplares

DIAGRAMACIÓN E IMPRESIÓN:

ARTES GRÁFICAS
Senefelder
Fundada en 1921

Agradecimientos

A mi madre, padre, hijo, esposo, hermanas, familia y amigos, por su apoyo constante.

Al equipo de profesores del Centro de Estudios de Turismo de la Universidad de Matanzas “Camilo Cienfuegos” (Cuba), por su ayuda y colaboración.

A la Universidad ECOTEC de Guayaquil, por su empeño en el lanzamiento del libro.

A todos, muchas gracias.

Dedicatoria

A mi segundo padre, **Argelio Frías**, por ser el gran inspirador de esta obra e impulsor constante de mi labor científico-investigativa.

Prólogo

La autora de esta obra me ha pedido de manera generosa escribirle un prólogo. No podía negarme a ello y las razones son varias: desde el afecto personal, la admiración profunda a su audacia, creatividad y valentía científica, hasta cierto compromiso que nace de haberla incitado a escribir un texto que, por una parte, facilitara la instrumentación e integración del arsenal estadístico a la investigación científica de los estudiantes de la carrera de Licenciatura en Turismo y, de otra, a conseguir ese mismo objetivo, pero ya en el ámbito del postgrado, concisamente en el desarrollo de las maestrías y doctorados así como de los variados proyectos que en ese campo ejecutan infinidad de profesionales.

El reto consistía entonces en escribir un texto no de la estadística en sí ni para sí -cuestión muy común en el modelo didáctico actual de la mayoría de los que se dedican a explicar esta materia y que es la causa fundamental del fracaso de aquellos que la cursan al resultarle tan sumamente abstracta y poco viable en sus aplicaciones, que lo que provoca en ellos es el rechazo- sino de exponer de forma amena y accesible un herramental, complejo en sí mismo, pero con un grado de utilidad insustituible en la investigación científica turística. Se trataba de exponer y explicar la ciencia de la estadística no como un fin en sí misma, sino como un medio, como un instrumento entendible y necesario para penetrar en la esencia de los fenómenos y apropiarse creadoramente de la materia investigada. Había que vencer, además, otro obstáculo y era llevar o extrapolar de manera creativa, los enfoques ancestrales probados en la manufactura, en el mundo de los tangibles, al mundo de los servicios, al campo de los intangibles, donde por cierto, las distribuciones no son nada normales, sino las más anormales con que se enfrenta la inteligencia humana y en este caso, demostrar la validez y pertinencia de los mismos principios de esta ciencia.

¿Cómo ha logrado resolver la autora estos dos grandes retos? A mi juicio, porque ha sabido conjugar de manera adecuada el método de la investigación con el método de la exposición. El primero le ha permitido en un tiempo relativamente corto, asimilar de manera profunda y en sus detalles la materia estudiada, el segundo, seguir una lógica sencilla para desmontar lo que el viejo modelo didáctico hacía extremadamente complejo y hacerlo asimilable e incluso, motivante a su público objetivo.

Varias razones explican el porqué la autora ha conseguido realizar de manera exitosa este proceso simbiótico, no dado a todos los que se arriesgan a escribir cosas para que otros lean y utilicen. Destacaría algunas, que a mi modo de ver, pudieran servir como patrón de comportamiento a otros.

Es una amante desenfrenada de la polémica, inveterada cuestionadora de "lo establecido" y un artífice de los "por qué", hurgando hasta en los más mínimos detalles para solo después emitir un juicio. No puedo dejar de mencionar su eficaz y eficiente proceso de formación, desde los primeros años de la carrera, combinando lo académico con lo científico-investigativo-laboral en todo el sistema empresarial turístico cubano: las agencias de viajes receptivas, incluidas sus prácticas en las actividades aeroportuarias, los sistemas de transportación turística, los grupos y cadenas hoteleras nacionales y extranjeras que operan en Cuba, su paso fecundo por todas las modalidades y tipos de productos turísticos que conforman el destino Cuba, entre otras. El haber realizado el ejercicio de la docencia en el campo de la estadística, la gestión de calidad, la dirección integrada de proyectos, la operación hotelera, el tráfico aéreo, la contabilidad hotelera, la teoría del turismo, la planificación de ferias y convenciones, la seguridad e higiene en hoteles, la recepción hotelera, la planificación estratégica, los sistemas informáticos de reservas así como las matemáticas… explica su prolífera producción científica y su afán de aprender de manera constante. Todo ello la ha llevado a ofrecernos este producto.

El método de exposición empleado por la autora, basado en exponer lo esencial de la teoría para luego y a través de la ejemplificación realizar un acercamiento a la práctica, se enriquece con la propuesta de solución de los problemas diseñados mediante el uso de paquetes estadístico-informáticos profesionales. Esta sencilla modelación, supera con creces el punto de vista tradicional, pues constituye en sí mismo un acercamiento activo al problema de la "reducción de la variabilidad" como la mejor vía de la mejora de los procesos, de los productos y de los servicios, pero de una manera más eficaz y más eficiente, dando, además, un valor agregado a los profesionales relacionados con los negocios al crear en ellos un pensamiento estadístico superior que los pone en mejores condiciones para enfrentar el proceso de toma de decisiones y de controlar y reducir la variabilidad indicada, lo que como se sabe, conlleva al proceso de mejora. A mi juicio, esta es la mayor contribución que este libro realiza, es decir, aportar elementos que enriquecen el proceso de formación de lo que se ha dado en llamar "pensamiento estadístico", como un conjunto de principios y valores que permiten identificar procesos, caracterizarlos, cuantificarlos, controlar y reducir su variabilidad para implementar acciones de mejora por parte de los decisores ([1]).

Enhorabuena por este regalo que de seguro encontrará favorable acogida en todos aquellos que necesiten, quieran y deseen hacer "ciencia" sin complicaciones.

Dr.C. Roberto Argelio Frías Jiménez
Decano de la Facultad de Ciencias Económicas e Informática
Universidad de Matanzas "Camilo Cienfuegos"
CUBA

---

1 Snee, R.D. (1990). "Statitiscal Thinking and its Contributions to Total Quality", The American Statistician, 44: 116-121.
_____(1993). "Waht`s Missing in Statistical Education", The American Statistician, 47: 149-154.
_____(1999). "Discussion: Development and use of satistical thinking: A new era", International Statistical Review, 67 (3): 225-258.

Contenido

Introducción

El libro que a continuación se presenta, aborda variadas herramientas estadísticas aplicadas en el ámbito de los estudios o investigaciones turísticas. Con los avances de las tecnologías informáticas, cada una de las técnicas aquí abordadas, son llevadas a cabo mediante paquetes estadísticos tales como: el STATGRAPHICS Plus y el programa profesional SPSS.

Este libro ha sido concebido como texto docente complementario de la asignatura Estadística que se imparte en la mayoría de las carreras de Hotelería y Turismo. También para los profesionales, puede ser objeto de consulta durante cursos de postgrados, diplomados, maestrías y en su propio desempeño dentro del puesto de trabajo. Destáquese que constituye material valioso de apoyo para los maestrantes de la Maestría en Gestión Turística que ofrece el Centro de Estudios de Turismo de la Universidad de Matanzas "Camilo Cienfuegos" (Cuba), y también para el resto de los maestrantes del Ecuador cuya especialización está enfocada al sector turístico y hotelero.

Esta obra posee como objetivo fundamental, mostrar cómo se emplea cada una de las herramientas estadísticas de mayor utilización en la investigación turística, mediante el procesamiento con paquetes o programas informáticos, por tanto, pretende viabilizar la obtención de los resultados para su posterior análisis, evitando los extensos cálculos y fórmulas que se llevan a cabo a mano alzada.

El libro está constituido por once capítulos que engloban gran cantidad de herramientas estadísticas, ejemplificadas todas, en el ámbito turístico y hotelero. Al final de cada capítulo, se propone una ejercitación de lo aprendido y se ofrece, además, la solución al ejercicio, para que el estudiante pueda comprobar si llegó a dominar el empleo de la herramienta estadística mediante los paquetes informáticos.

El **capítulo 1** se dedica a abordar algunas generalidades relacionadas con la estadística como rama de las matemáticas, y ofrece datos acerca de los dos programas informáticos utilizados.

El **capítulo 2** se adentra en la estadística descriptiva, abordando las distribuciones de frecuencias.

El **capítulo 3** expone los diferentes estadígrafos o estadísticos descriptivos más empleados en las investigaciones, así como los gráficos asociados a los mismos.

El **capítulo 4** trata acerca de las dócimas o pruebas de hipótesis paramétricas, mientras que el **capítulo 5**, aborda las dócimas de hipótesis no paramétricas.

El **capítulo 6** hace énfasis en el análisis de varianza, en específico, el unifactorial.

El **capítulo 7** destaca los elementos más importantes a dominar del análisis de asociación.

El **capítulo 8** expone las particularidades del análisis factorial como parte de la estadística multivariante y, de la misma forma, el **capítulo 9** exhibe el tema relacionado con el análisis discriminante.

El **capítulo 10** aborda el análisis cluster o de conglomerados como parte de la estadística multivariante igualmente.

El **capítulo 11** y último, muestra cómo confeccionar o elaborar un diagrama de Pareto, así como algunos de los diferentes gráficos o cartas de control.

Se espera, finalmente, que el libro contribuya a elevar el pensamiento estadístico de aquellos que, más tarde, formen parte de los directivos que toman decisiones constantemente en el sector turístico y de los que actualmente, constituyen la gerencia de instalaciones turísticas y hoteleras.

Capítulo
1

Generalidades.

1.1. ¿Qué es la Estadística?

La Estadística es una rama de las matemáticas que se ocupa de reunir, organizar y analizar datos numéricos y que ayuda a resolver problemas como el diseño de experimentos y la toma de decisiones.

Tradicionalmente la estadística-matemática se divide en dos partes: la estadística descriptiva y la estadística inferencial. En la primera, se agrupan todas aquellas técnicas asociadas con el procesamiento de un conjunto de datos. En la segunda, se agrupan las que permiten la toma de decisiones mediante las conclusiones a que se arriben cuando se analizan características numéricas del fenómeno que se estudia.

1.2. Diferencias entre dato e información en la toma de decisiones.

Dato: es todo conjunto de caracteres que describe algo sobre nuestra realidad. En un Sistema de Información es el "input".

Información: es la parte de los datos que influye en las decisiones adoptadas, o que puede conducir a su modificación cuando no está disponible. En un Sistema de Información es el "output".

La diferencia entre dato e información no reside en el contenido del conjunto de caracteres, sino en la relación de éstos con el tipo de decisión. Una serie de caracteres puede ser datos para un decisor e información para otro. La información reduce la incertidumbre del decisor y su definición, solamente puede tener lugar dentro del marco de la toma de decisiones.

En el sector turístico, los empresarios o directivos toman diariamente miles de decisiones como parte esencial de su trabajo, por lo cual requieren tener disponible un conjunto de datos, de muchas naturalezas diferentes. Empleando diversas técnicas estadísticas, dichos datos ofrecen información valiosa que posibilita la acertada toma de decisiones. Todo ello está en consonancia con una frase que expresara hace poco más de un siglo H. G. Wells: *"el pensamiento estadístico un día será de tanta*

importancia y necesario como la capacidad de leer y escribir".

1.3. Tipos de datos u observaciones.

En general, los datos u observaciones recopilados por un investigador acerca del comportamiento de una variable, pueden ser de dos tipos: cualitativos o cuantitativos. Los datos cualitativos son aquellos que reflejan cualidades, por ejemplo: el estado civil de los trabajadores en un centro laboral, la categoría docente de los profesores de un departamento, etc. Los datos cuantitativos son aquellos que reflejan cantidades y es importante señalar dos tipos: los discretos y los continuos.

Los datos cuantitativos discretos sólo pueden tomar un número finito o numerable de valores enteros, por ejemplo: el número de estudiantes en un aula, el número de ausencias de un trabajador en un mes, el número de habitaciones disponibles en un hotel, la cantidad de piezas producidas por una fábrica en una semana, etc. Los datos cuantitativos continuos, en cambio, pueden tomar un número infinito de valores reales, por ejemplo: la estatura de una persona, la edad de varias personas, el rendimiento de un determinado cultivo en un huerto, el tiempo de demora de un estudiante al responder un examen, etc.

1.4. ¿A qué se le denomina "parámetro" y a qué se le llama "variable"?

Parámetro: constituye una característica poblacional que se desea investigar y suele ser desconocida a priori. Cuando la característica es numérica (o sea, se puede medir), se denomina variable. Cuando la característica no puede ser medida numéricamente, se denomina atributo.

Las variables pueden ser discretas o continuas. La mayor parte de las variables continuas pueden tratarse como discretas.
Los atributos presentan categorías y pueden clasificarse como ordenables y no ordenables.

Las variables y los atributos presentan 4 tipos de escalas de medición:
- nominal o clasificatoria
- ordinal
- intervalo
- razón o proporción

1.5. Paquete informático de procesamiento estadístico SPSS.

El **SPSS** *(**S**tatistical **P**roduct for **S**ervice **S**olutions*) es un programa estadístico informático muy usado en las ciencias sociales y empresas de investigación de mercado. Originalmente SPSS era el acrónimo de "Statistical Package for the Social Sciences". En la actualidad, la sigla designa tanto el programa como la empresa que

lo produce.

Dicho programa fue creado en 1968 por Norman H. Nie, C. Hadlai Hull y Dale H. Bent. Entre 1969 y 1975 la Universidad de Chicago por medio de su National Opinion Research Center, estuvo a cargo del desarrollo, distribución y venta del programa. A partir de 1975 corresponde a SPSS Inc.

Como programa estadístico, es muy popular su uso, debido a la capacidad de trabajar con bases de datos de gran tamaño y por permitir, además, la recodificación de las variables y registros según las necesidades del usuario. El programa consiste en un módulo base y módulos anexos que se han ido actualizando constantemente con nuevos procedimientos estadísticos.

En las dócimas de hipótesis (paramétricas y no paramétricas), el investigador tiene que dominar cuáles hipótesis corresponden con el caso de estudio, y decidir con qué nivel de significación va a trabajar. El resto, lo realiza el software estadístico.

1.6. Paquete informático de procesamiento estadístico STATGRAPHICS Plus.

El STATGRAPHICS Plus es un software práctico que permite concentrarse en los conceptos y resultados estadísticos, sin tener que malgastar tiempo aprendiendo sofisticados programas. Su nombre está conformado por “STAT” de “statistic” (estadística) y “GRAPHICS” de gráficos. Es un programa de estadística cuyo fabricante es Manugistics, y por áreas de aplicación, se identifican los siguientes campos: estadística descriptiva, calidad, etc.

El programa presenta un importante conjunto de novedades, que aumentan sus reconocidas capacidades en cuanto a potencia de cálculo y gráfica, flexibilidad, racionalidad, facilidad de uso y relación prestaciones/precio. Las principales novedades son las siguientes:

- un Editor Estadístico (StatReport) que desde el propio STATGRAPHICS Plus, ofrece la posibilidad de preparar informes con gráficos y tablas cambiantes cuando cambian los correspondientes datos y análisis efectuados. Este editor permite cambiar tipos de letra, color y tamaño de las mismas, así como configurar páginas, sin salir de STATGRAPHICS Plus
- un Asistente Estadístico (StatWizard), que ayuda de una manera efectiva a seleccionar, en cada caso, el procedimiento que mejor se adecue para recopilar y analizar los datos
- un Enlace Estadístico (StatLink), que permite enlazar el Libro de Análisis (Statfolio) con el que se esté trabajando, con la fuente de datos
- mayor y más fácil control de las reconocidas Capacidades Gráficas del paquete, a través de un solo cuadro de diálogo

Capítulo
2

Distribuciones de frecuencias.

2.1. Tipos de distribuciones de frecuencias unidimensionales.

Cuando un investigador recopila un conjunto de datos u observaciones acerca de una variable determinada, la primera tarea consiste en ordenarlos, para luego, poder extraer de ellos toda la información posible que viabilice la toma posterior de decisiones. Para ello, una tabla de distribución de frecuencias resulta de mucha utilidad.

Existen 3 tipos de distribuciones de frecuencias unidimensionales:
- tipo I: los valores no se repiten en ningún caso
- tipo II: cada valor de la característica medida se repite un determinado número de veces
- tipo III: las observaciones están clasificadas en intervalos

2.2. Tipos de distribuciones de frecuencias bidimensionales.

Estas distribuciones aparecen cuando de una población se desean estudiar dos variables.

Existen 3 tipos de distribuciones de frecuencias bidimensionales:
- cuando las dos informaciones son atributos
- cuando una información corresponde a una variable y la otra a un atributo
- cuando las dos informaciones son variables

Cuando las variables son cuantitativas, a las tablas de frecuencias se les denomina "tablas de correlación", y cuando se trata de atributos o variables cualitativas, se les llama "tablas de contingencia".

El gráfico más utilizado en las distribuciones de frecuencias bidimensionales, es el gráfico de dispersión.

2.3. Tablas de distribución de frecuencias unidimensionales.

Cada tipo de distribución de frecuencias, puede representarse en una tabla, que resulta de mucha utilidad para ordenar los datos y extraer la información precisa.

La tabla está compuesta por varias columnas. La primera representa las clases (X) o diferentes valores que toma la variable a analizar. La segunda representa la frecuencia absoluta (n_i) que recoge la cantidad de veces que aparece cada clase. La tercera, representa la frecuencia relativa (f_i) o porcentaje de cada clase. La cuarta representa la frecuencia absoluta acumulada (N_i) de cada clase que consiste en la suma de las frecuencias absolutas de las clases anteriores. Por último, la quinta columna representa la frecuencia relativa acumulada (F_i) de cada clase, que consiste en la suma de las frecuencias relativas de las clases anteriores.

Véase un ejemplo.

Ejemplo 1:

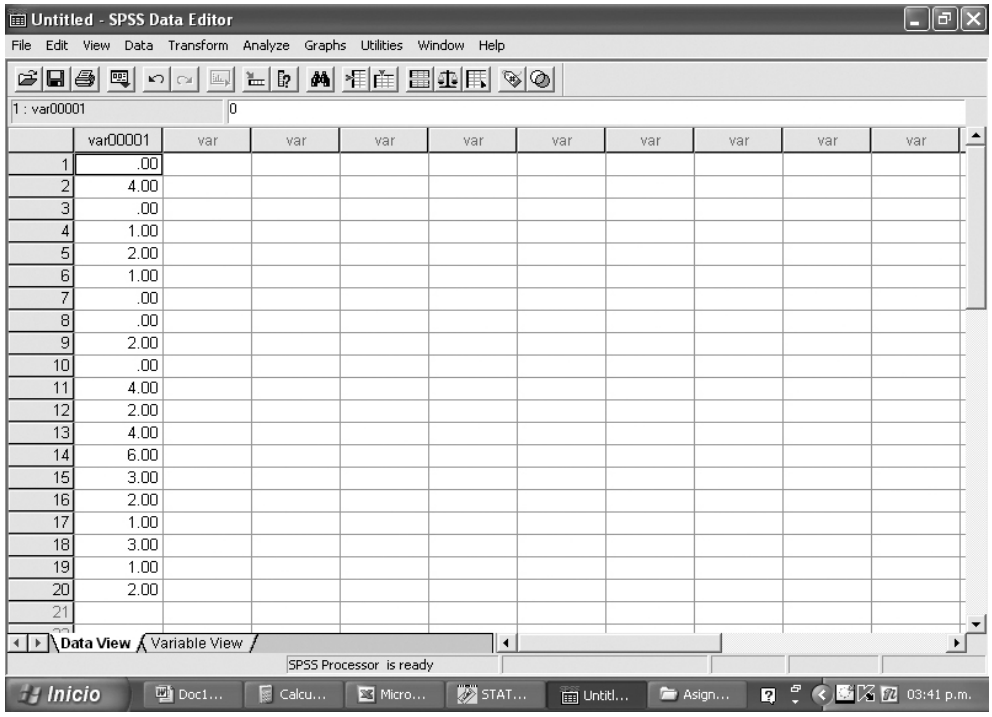
El Departamento de Recursos Humanos del Hotel Z, ha recopilado la cantidad de ausencias de los 20 trabajadores que laboran en la lavandería de la entidad, durante el mes de marzo de 2009:

0	2	2	4	1
4	1	0	6	3
0	0	4	3	1
1	0	2	2	2

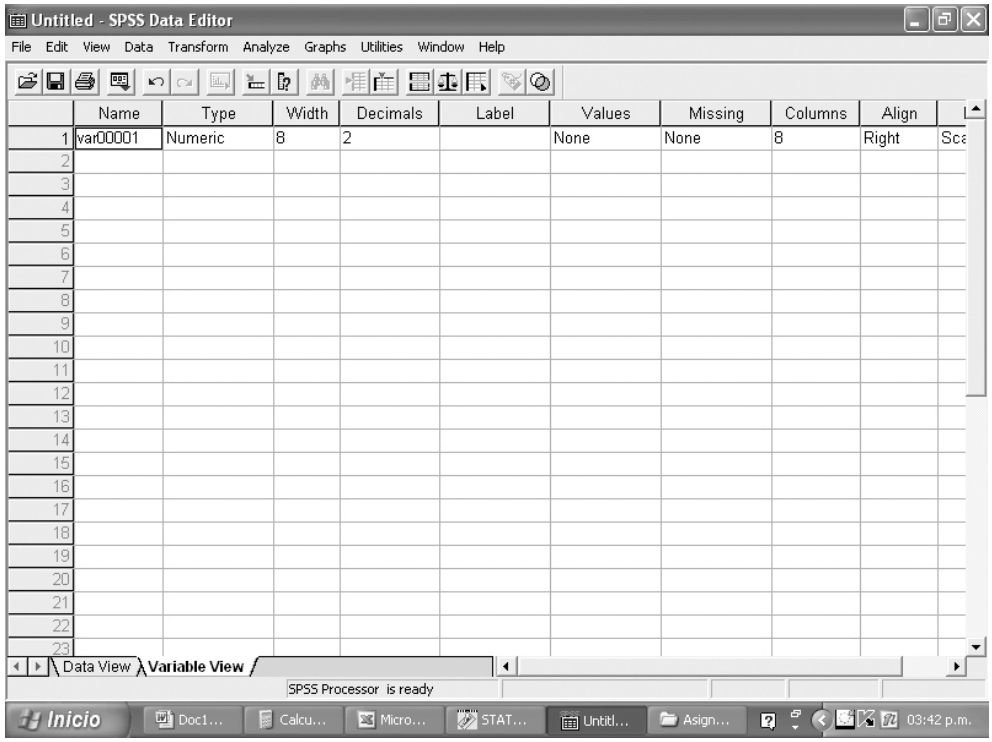
Solución:

Variable cuantitativa discreta: cantidad de ausencias de un grupo de trabajadores.

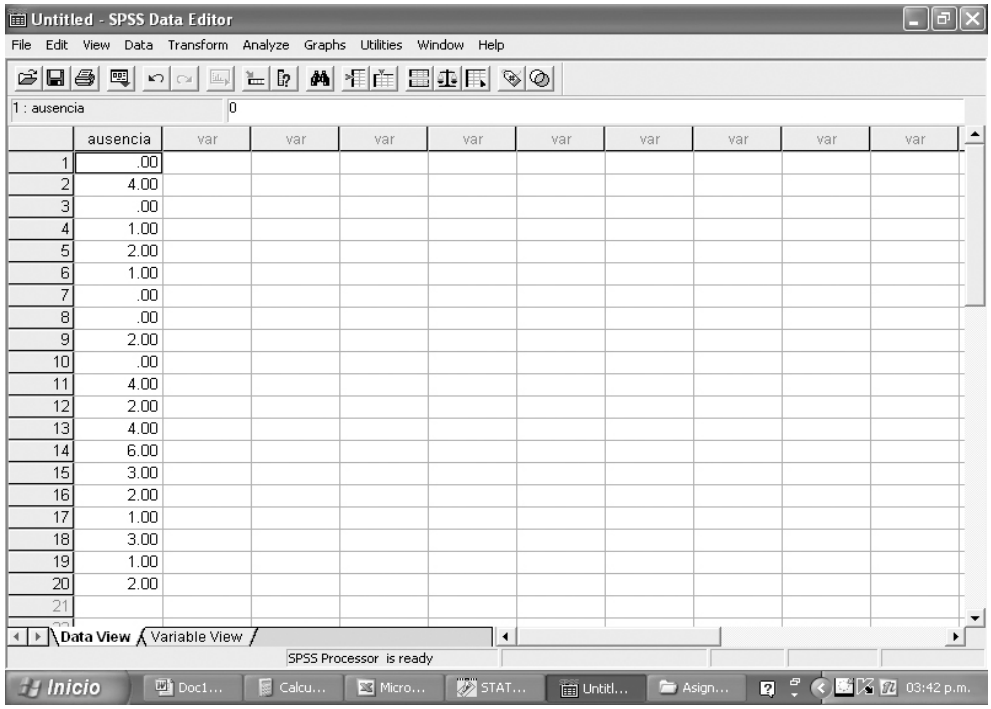
Para confeccionar la tabla de distribución de frecuencias anterior utilizando el SPSS, primeramente resulta imprescindible saber que este software estadístico reconoce que cada columna de datos, constituye una variable de análisis. Por tanto, en este sencillo ejemplo, se recordará que sólo se tiene una variable (cantidad de ausencias). Habrá entonces sólo una columna de observaciones en el SPSS en la vista de datos (data view), y en la vista de variable (variable view) se colocará el nombre de la variable.



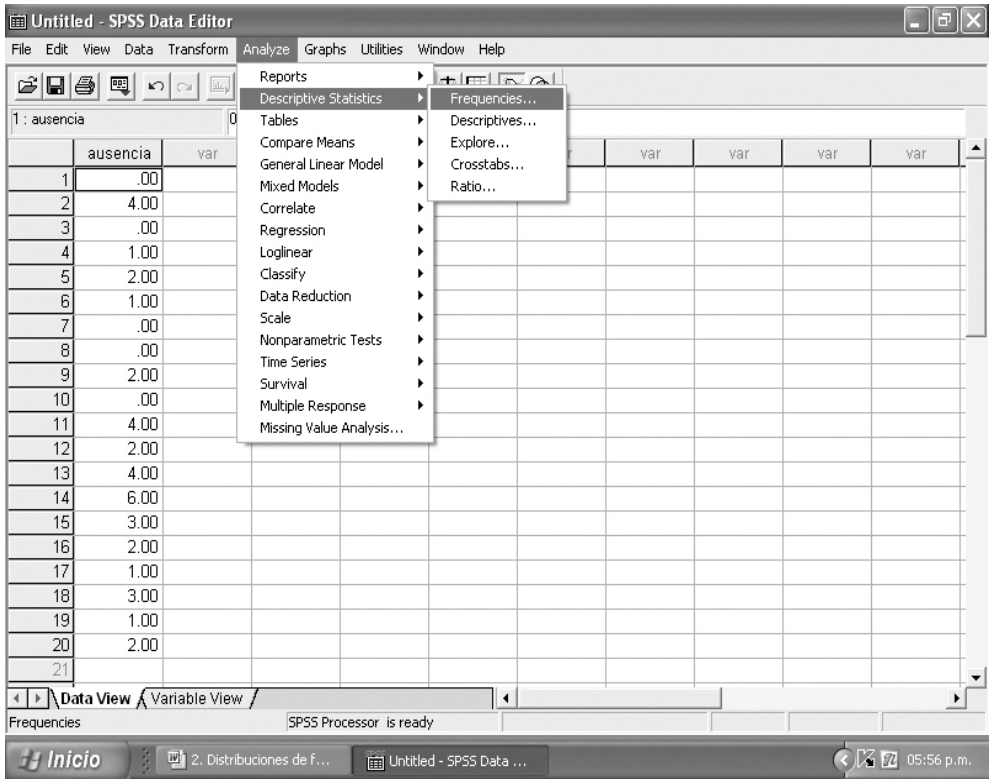
Vista de datos.

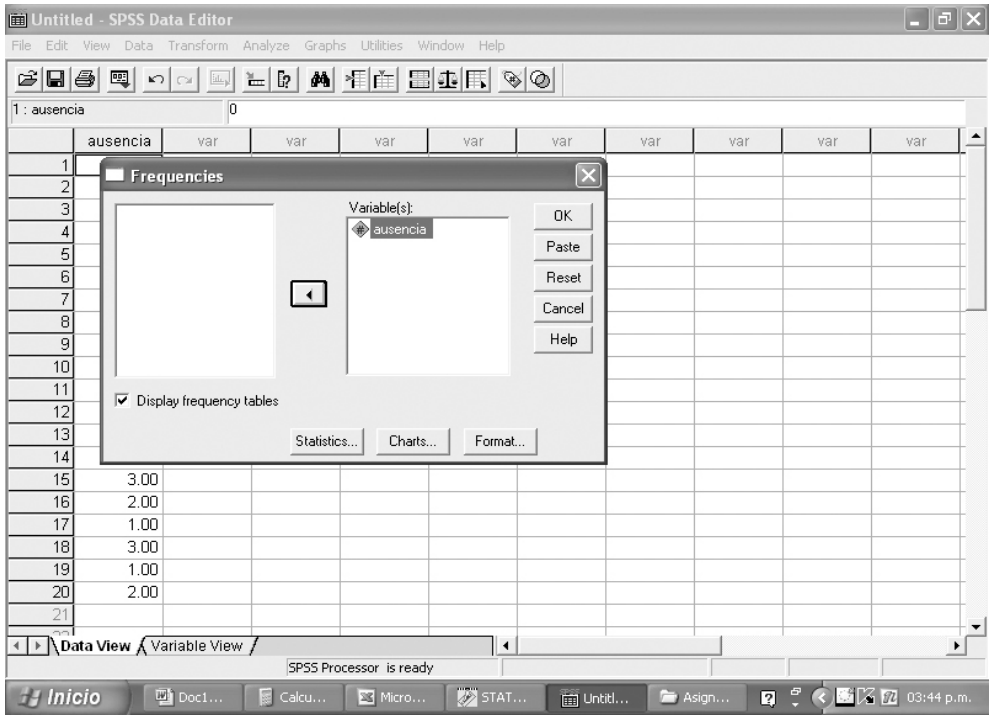


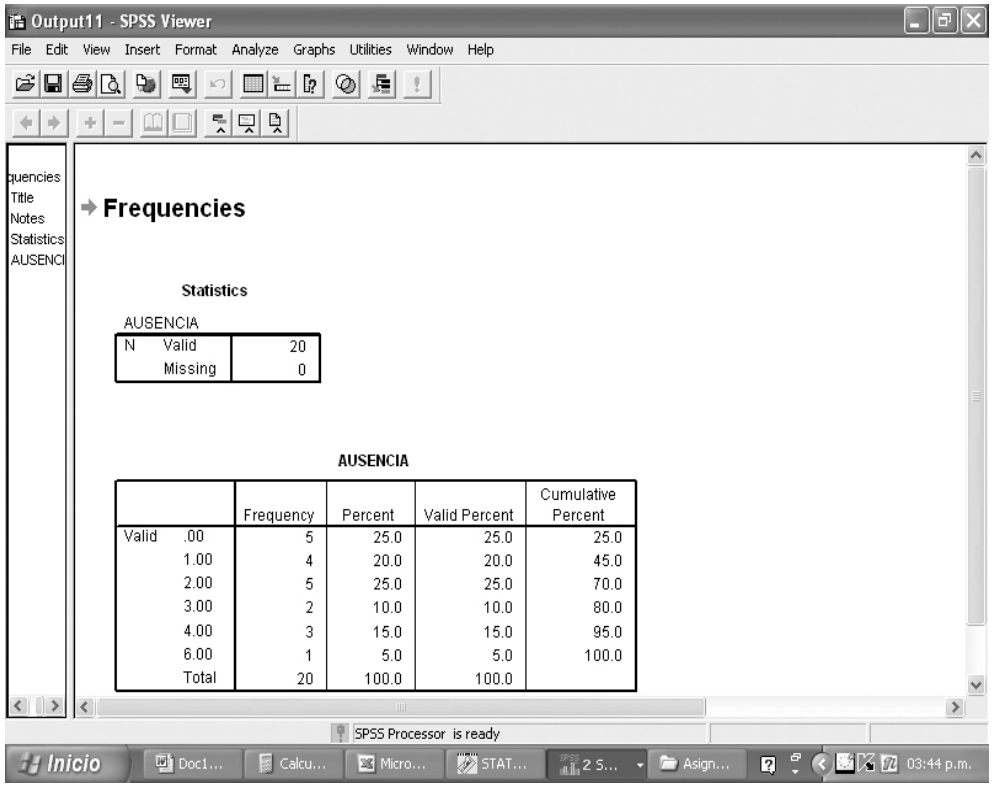
Vista de variable.



Datos colocados con el nombre de la variable.




Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : ausencia
0
ausencia
var
Frequencies
Variable(s):
ausencia
OK
Paste
Reset
Cancel
Help
Display frequency tables
Statistics...
Charts...
Format...
3.00
2.00
1.00
3.00
1.00
2.00
Data View
Variable View
SPSS Processor is ready
Inicio
03:44 p.m.


Output11 - SPSS Viewer
File Edit View Insert Format Analyze Graphs Utilities Window Help
quencies
Title
Notes
Statistics
AUSENCI
Frequencies
Statistics
AUSENCIA
N Valid 20
Missing 0
AUSENCIA
Frequency Percent Valid Percent Cumulative Percent
Valid .00 5 25.0 25.0 25.0
1.00 4 20.0 20.0 45.0
2.00 5 25.0 25.0 70.0
3.00 2 10.0 10.0 80.0
4.00 3 15.0 15.0 95.0
6.00 1 5.0 5.0 100.0
Total 20 100.0 100.0
SPSS Processor is ready
Inicio
03:44 p.m.

Los datos recopilados y ordenados en la tabla de distribución de frecuencias anterior, transmiten la siguiente información:

- cinco trabajadores no presentaron ninguna ausencia durante el mes
- dos trabajadores tuvieron 3 ausencias
- un 20% del total tuvo 4 ausencias
- el 80% de los trabajadores tuvo hasta 3 ausencias
- el total de empleados analizados es de 20

Véase otro ejemplo.

Ejemplo 2:

El Departamento de Comercial de la Agencia de Viajes Z, ha recopilado las edades de un grupo de 15 clientes canadienses que compraron la excursión "Habana Colonial" en el día de ayer. Los datos son los siguientes:

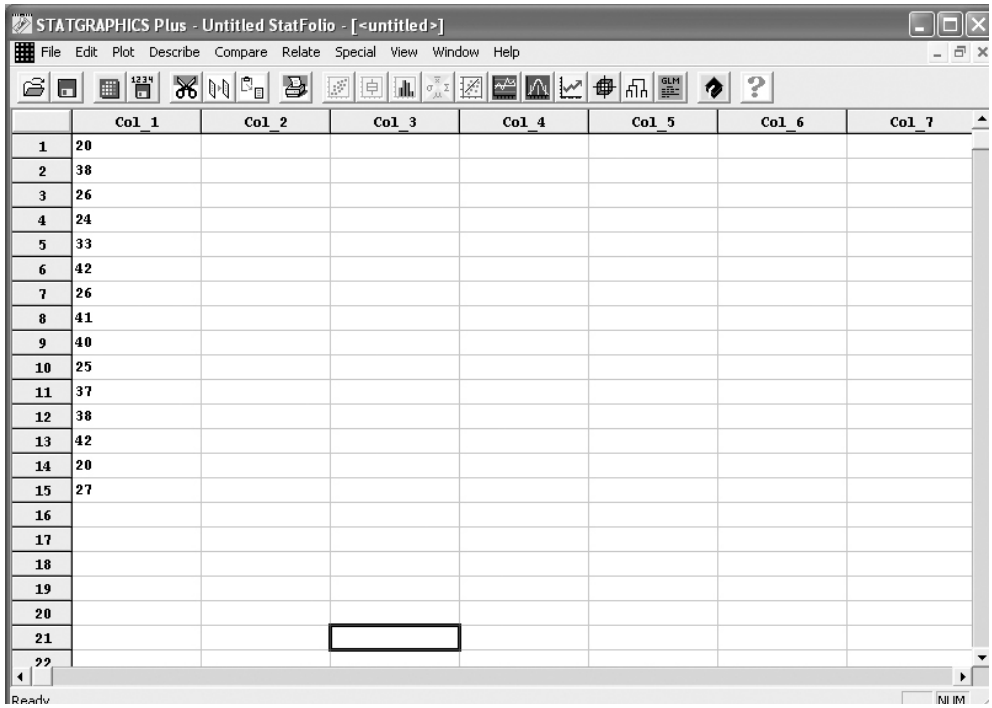
20	38	26
24	33	42
26	41	40
25	37	38
42	20	27

Solución:

Variable cuantitativa continua: edad de un grupo de clientes.

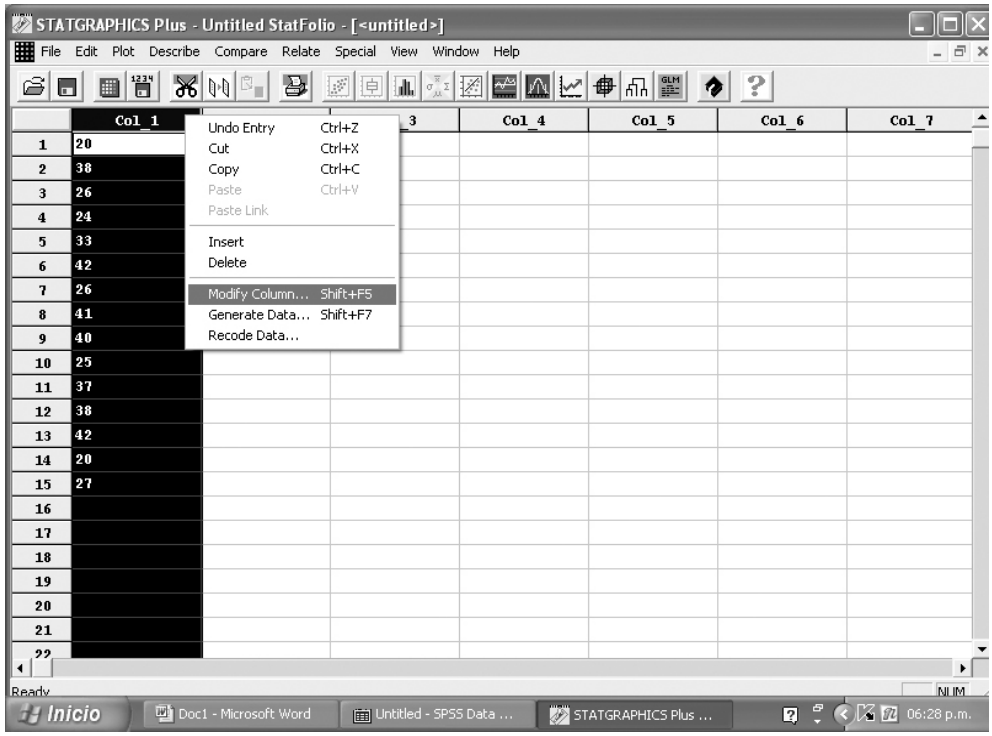
Como las variables continuas pueden tomar muchos valores, no sólo enteros sino también decimales, estos tienden a agruparse en intervalos de clases. La cantidad de intervalos de clases (K) debe oscilar entre 5 y 20.

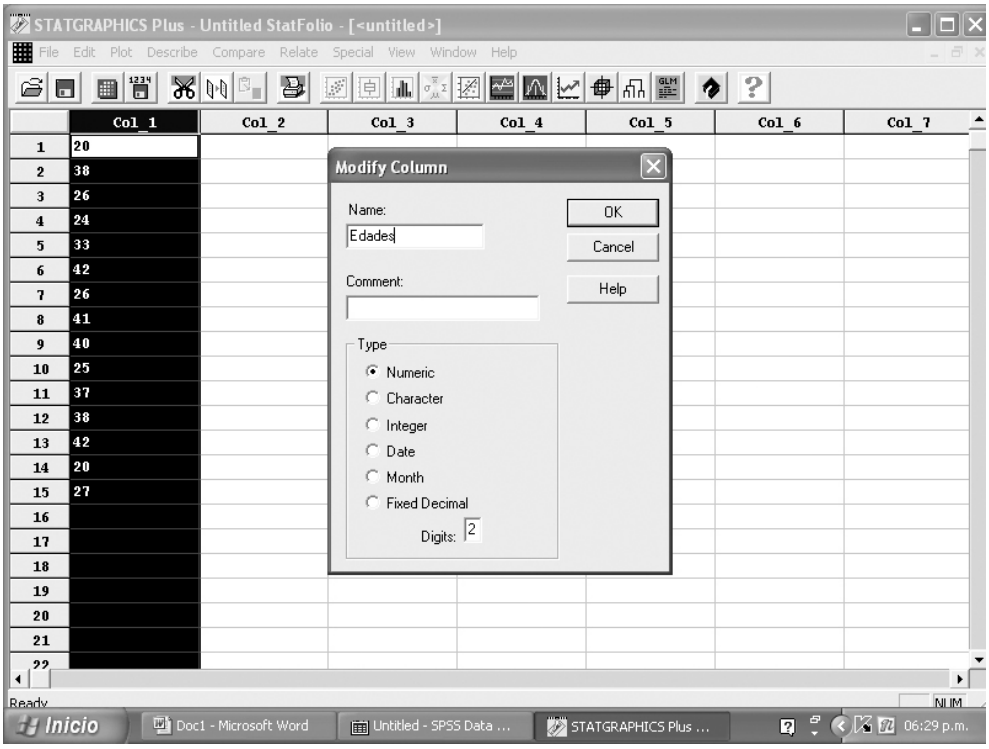
Para confeccionar la tabla utilizando el software estadístico STATGRAPHICS Plus, se colocan los datos de la siguiente manera:

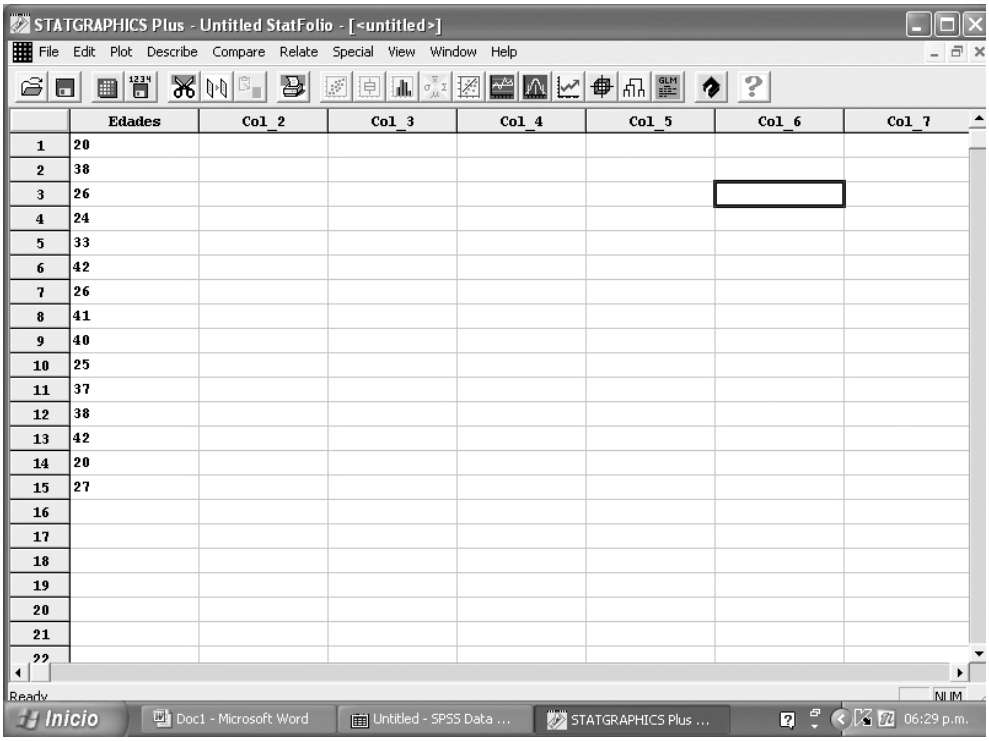


Vista de datos

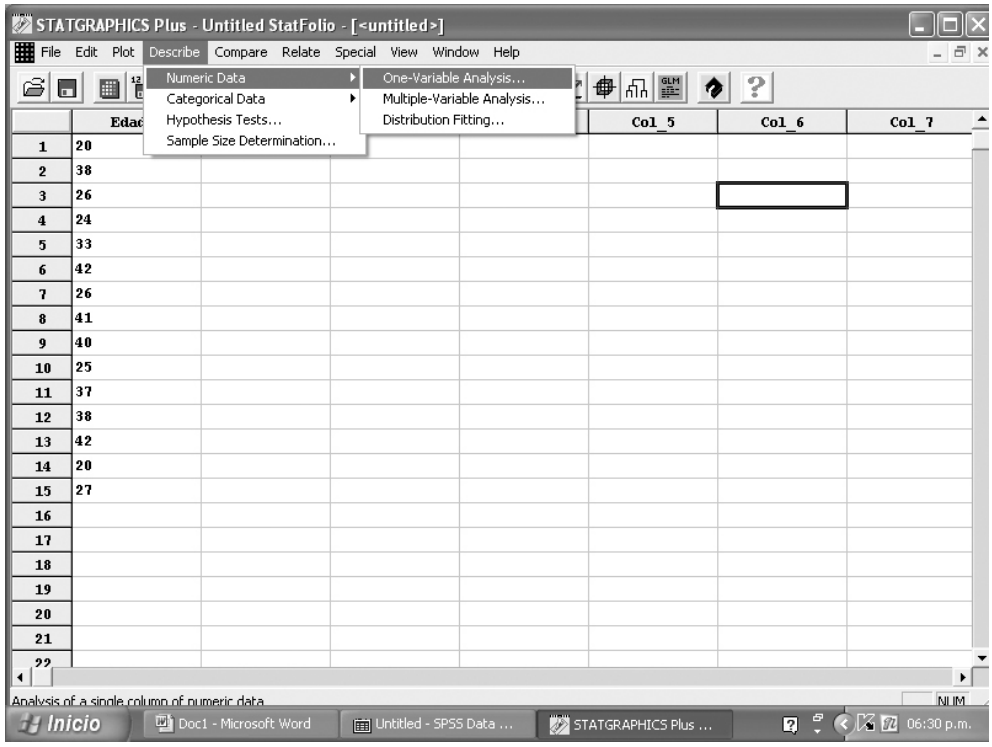
Para cambiarle el nombre a la columna que contiene los datos de la variable (edad), sería:

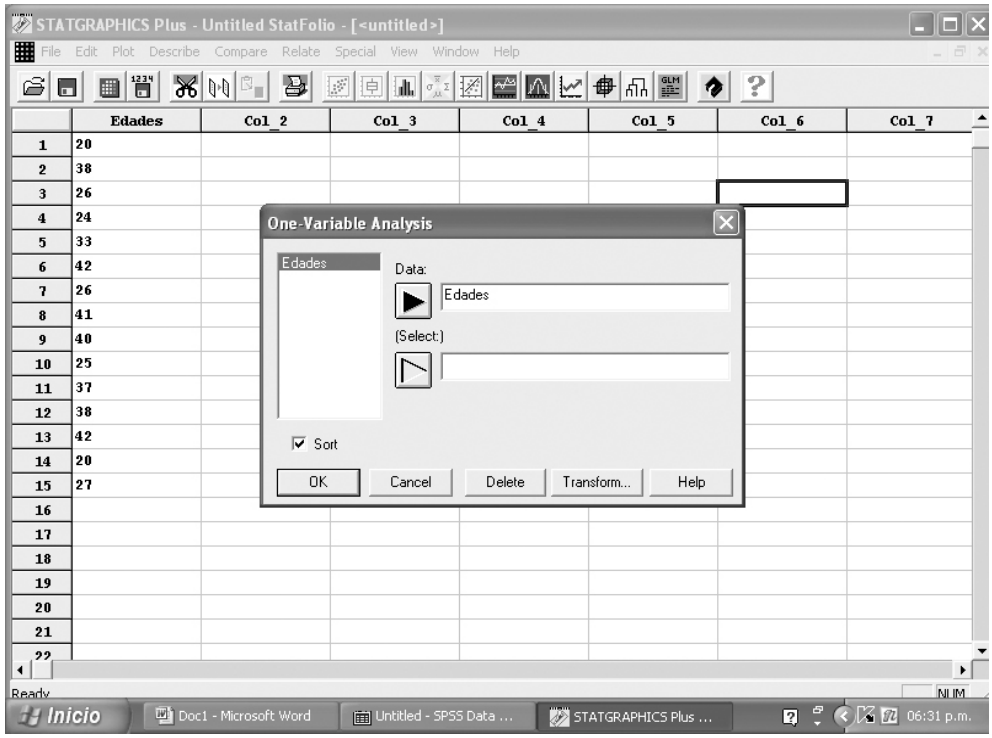


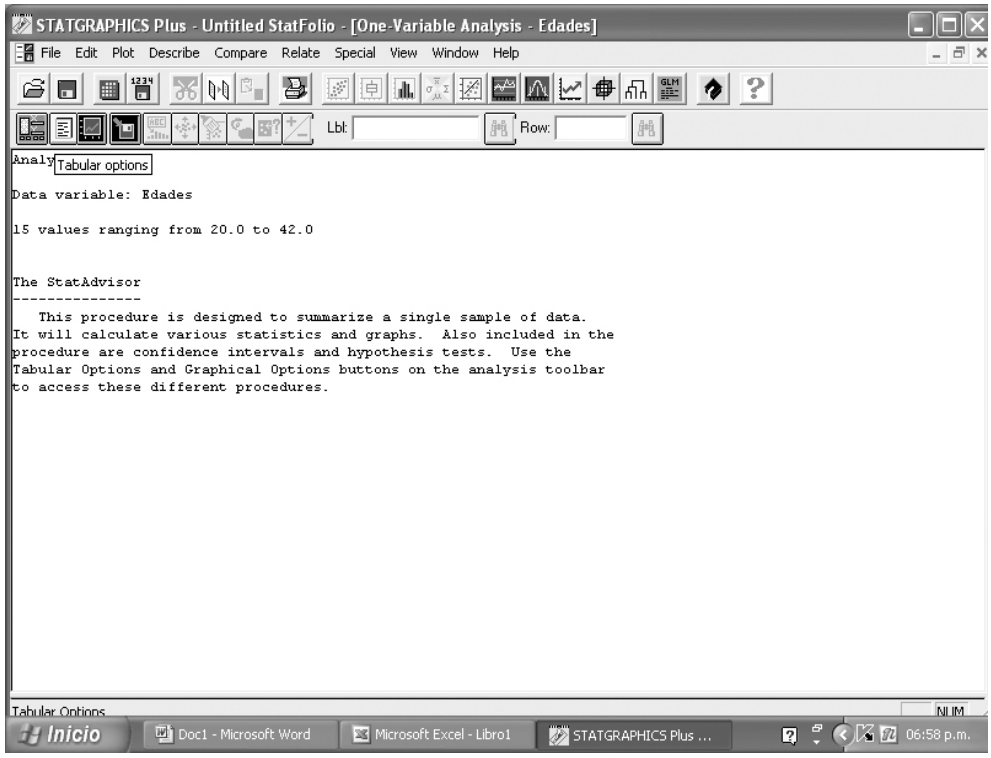


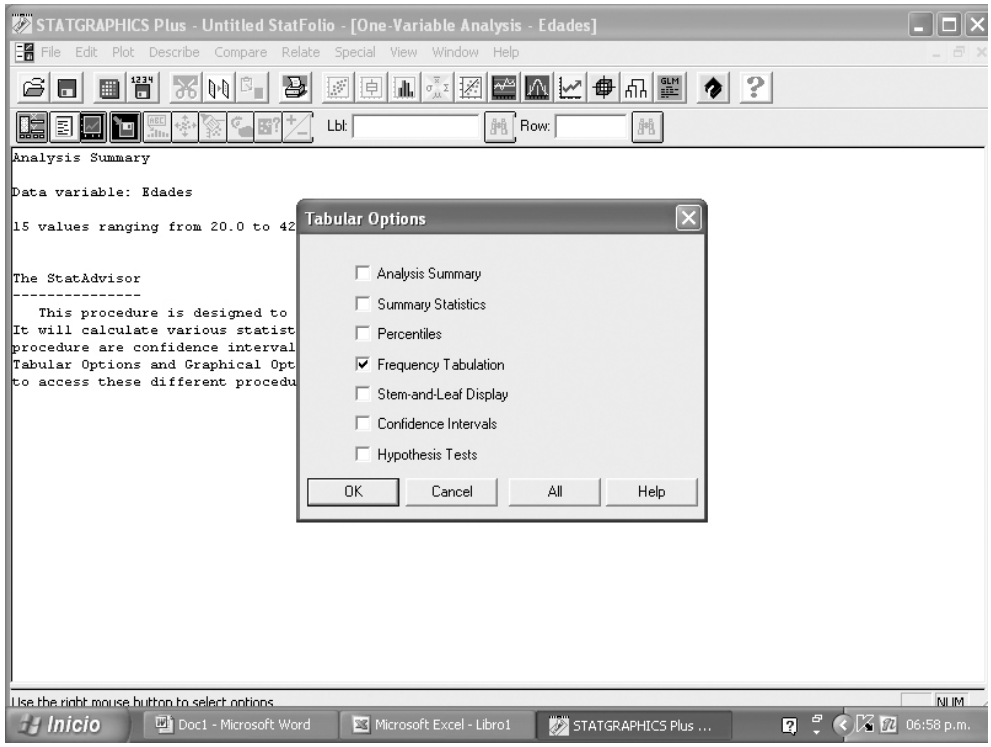


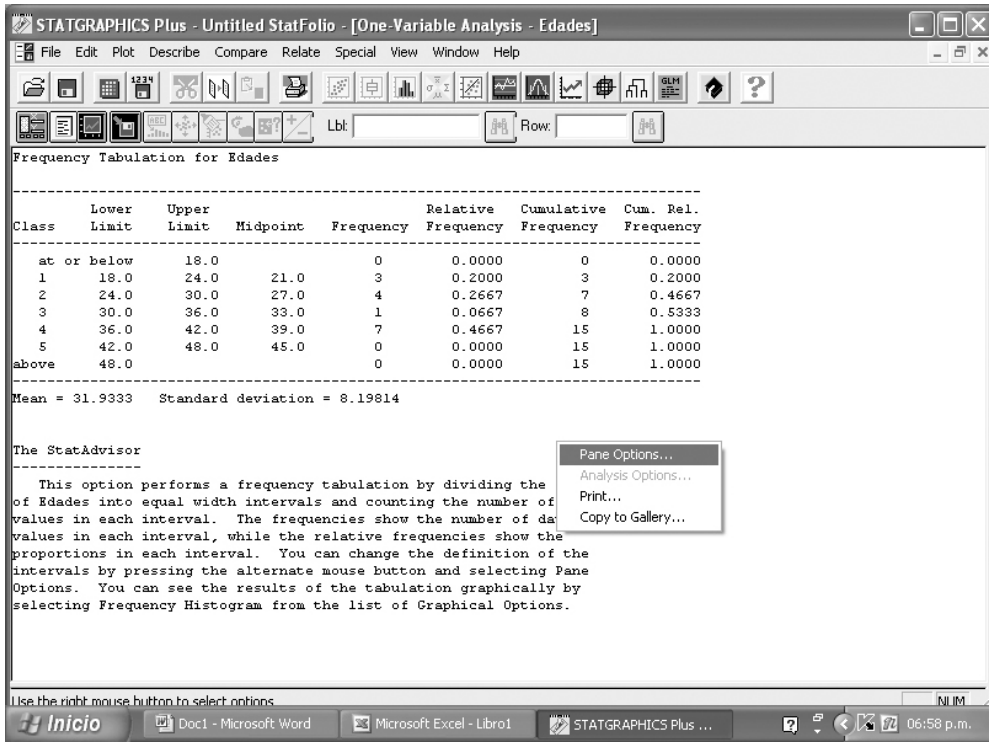
Vista de datos con el nombre de la variable.

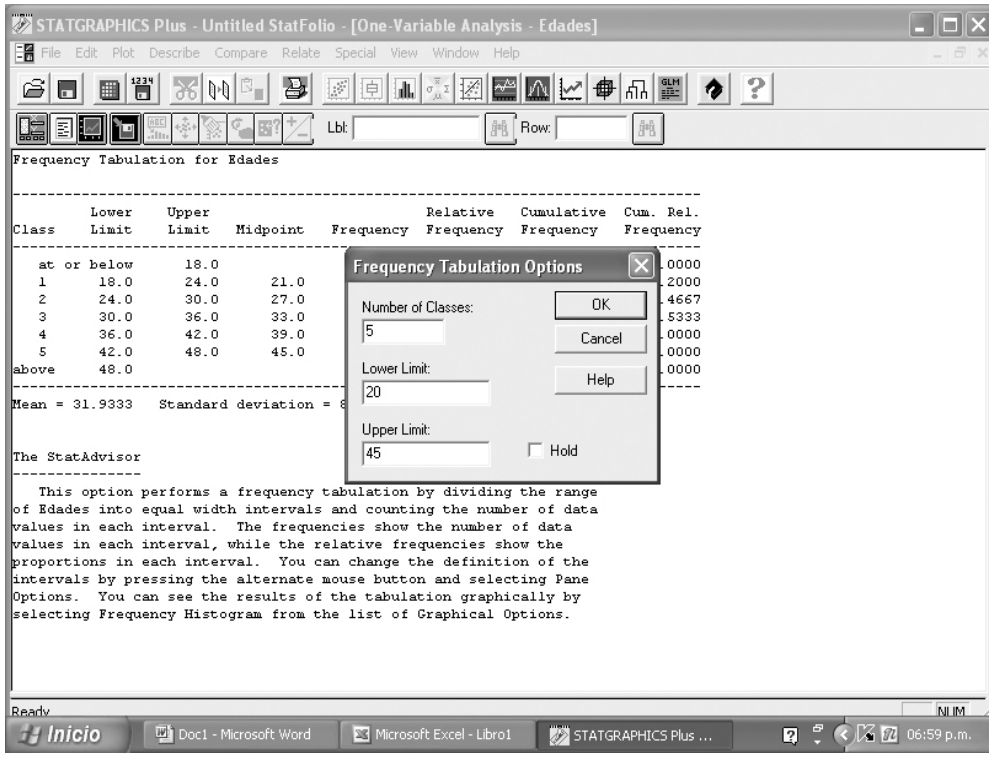

STATGRAPHICS Plus - Untitled StatFolio - [<untitled>]
File Edit Plot Describe Compare Relate Special View Window Help
Numeric Data
Categorical Data
Hypothesis Tests...
Sample Size Determination...
One-Variable Analysis...
Multiple-Variable Analysis...
Distribution Fitting...
Col_5 Col_6 Col_7
1 20
2 38
3 26
4 24
5 33
6 42
7 26
8 41
9 40
10 25
11 37
12 38
13 42
14 20
15 27
Analysis of a single column of numeric data
NUM
Inicio
Doc1 - Microsoft Word
Untitled - SPSS Data ...
STATGRAPHICS Plus ...
06:30 p.m.

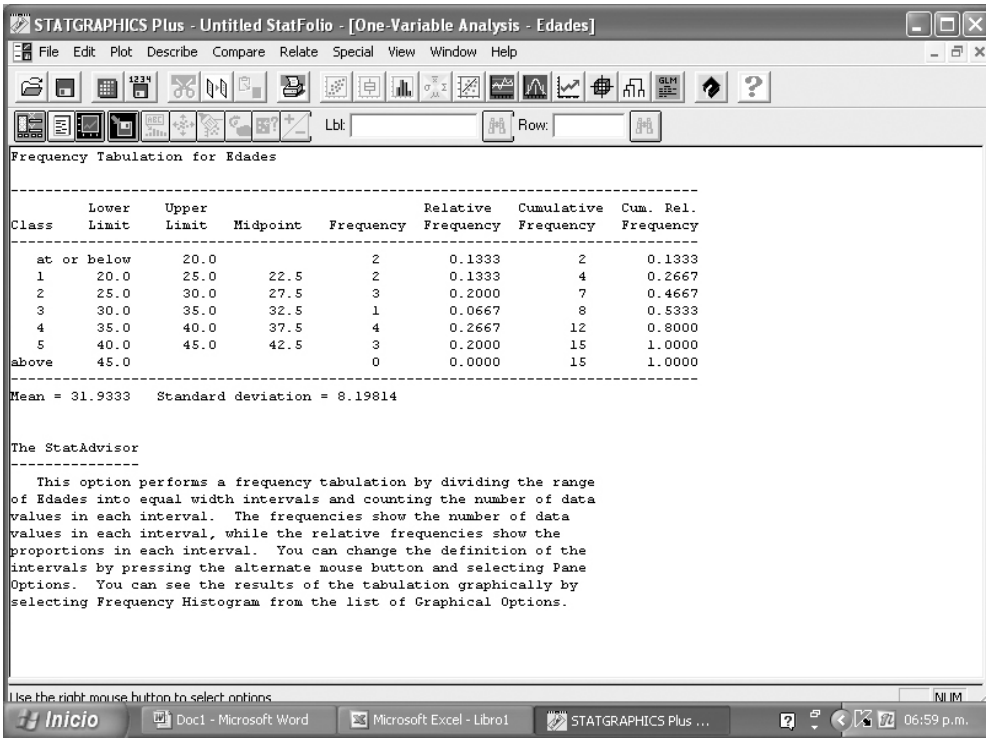

STATGRAPHICS Plus - Untitled StatFolio - [<untitled>]
File Edit Plot Describe Compare Relate Special View Window Help
Edades Col_2 Col_3 Col_4 Col_5 Col_6 Col_7
1 20
2 38
3 26
4 24
5 33
6 42
7 26
8 41
9 40
10 25
11 37
12 38
13 42
14 20
15 27
One-Variable Analysis
Edades
Data:
Edades
(Select:)
Sort
OK
Cancel
Delete
Transform...
Help
Ready
NUM
Inicio
Doc1 - Microsoft Word
Untitled - SPSS Data ...
STATGRAPHICS Plus ...
06:31 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [One-Variable Analysis - Edades]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Analy
Tabular options
Data variable: Edades
15 values ranging from 20.0 to 42.0
The StatAdvisor
This procedure is designed to summarize a single sample of data.
It will calculate various statistics and graphs. Also included in the
procedure are confidence intervals and hypothesis tests. Use the
Tabular Options and Graphical Options buttons on the analysis toolbar
to access these different procedures.
Tabular Options
NUM
Inicio
Doc1 - Microsoft Word
Microsoft Excel - Libro1
STATGRAPHICS Plus ...
06:58 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [One-Variable Analysis - Edades]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Analysis Summary
Data variable: Edades
15 values ranging from 20.0 to 42
The StatAdvisor
This procedure is designed to
It will calculate various statist
procedure are confidence interval
Tabular Options and Graphical Opt
to access these different procedu
Tabular Options
Analysis Summary
Summary Statistics
Percentiles
Frequency Tabulation
Stem-and-Leaf Display
Confidence Intervals
Hypothesis Tests
OK
Cancel
All
Help
Use the right mouse button to select options
NUM
Inicio
Doc1 - Microsoft Word
Microsoft Excel - Libro1
STATGRAPHICS Plus ...
06:58 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [One-Variable Analysis - Edades]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Frequency Tabulation for Edades
Class | Lower Limit | Upper Limit | Midpoint | Frequency | Relative Frequency | Cumulative Frequency | Cum. Rel. Frequency
at or below | | 18.0 | | 0 | 0.0000 | 0 | 0.0000
1 | 18.0 | 24.0 | 21.0 | 3 | 0.2000 | 3 | 0.2000
2 | 24.0 | 30.0 | 27.0 | 4 | 0.2667 | 7 | 0.4667
3 | 30.0 | 36.0 | 33.0 | 1 | 0.0667 | 8 | 0.5333
4 | 36.0 | 42.0 | 39.0 | 7 | 0.4667 | 15 | 1.0000
5 | 42.0 | 48.0 | 45.0 | 0 | 0.0000 | 15 | 1.0000
above | 48.0 | | | 0 | 0.0000 | 15 | 1.0000
Mean = 31.9333 Standard deviation = 8.19814
The StatAdvisor
This option performs a frequency tabulation by dividing the
of Edades into equal width intervals and counting the number of
values in each interval. The frequencies show the number of da
values in each interval, while the relative frequencies show the
proportions in each interval. You can change the definition of the
intervals by pressing the alternate mouse button and selecting Pane
Options. You can see the results of the tabulation graphically by
selecting Frequency Histogram from the list of Graphical Options.
Pane Options...
Analysis Options...
Print...
Copy to Gallery...
Use the right mouse button to select options
NUM
Inicio
Doc1 - Microsoft Word
Microsoft Excel - Libro1
STATGRAPHICS Plus ...
06:58 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [One-Variable Analysis - Edades]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Frequency Tabulation for Edades
Class | Lower Limit | Upper Limit | Midpoint | Frequency | Relative Frequency | Cumulative Frequency | Cum. Rel. Frequency
at or below | | 18.0 | | | | | .0000
1 | 18.0 | 24.0 | 21.0 | | | | .2000
2 | 24.0 | 30.0 | 27.0 | | | | .4667
3 | 30.0 | 36.0 | 33.0 | | | | .5333
4 | 36.0 | 42.0 | 39.0 | | | | .0000
5 | 42.0 | 48.0 | 45.0 | | | | .0000
above | 48.0 | | | | | | .0000
Mean = 31.9333 Standard deviation = 8
Frequency Tabulation Options
Number of Classes: 5
Lower Limit: 20
Upper Limit: 45
OK
Cancel
Help
Hold
The StatAdvisor
This option performs a frequency tabulation by dividing the range
of Edades into equal width intervals and counting the number of data
values in each interval. The frequencies show the number of data
values in each interval, while the relative frequencies show the
proportions in each interval. You can change the definition of the
intervals by pressing the alternate mouse button and selecting Pane
Options. You can see the results of the tabulation graphically by
selecting Frequency Histogram from the list of Graphical Options.
Ready
NUM
Inicio
Doc1 - Microsoft Word
Microsoft Excel - Libro1
STATGRAPHICS Plus ...
06:59 p.m.



La tabla muestra la siguiente información:

- dos clientes tienen hasta 20 años
- un solo cliente tiene entre 30 y 35 años de edad
- un 20% del total tiene entre 25 y 30 años
- doce clientes tienen hasta 40 años
- el 47% de los clientes tiene hasta 30 años de edad
- fueron estudiados 15 clientes

EJERCITACIÓN

El Departamento de Calidad y Atención al Cliente del Hotel W, ha recopilado la cantidad de quejas que un grupo de clientes canadienses, ha expresado durante su semana de estancia en la instalación. Los datos se muestran a continuación:

3	3	6
5	1	4
4	4	1
6	2	3

a) ¿Cuántos clientes fueron analizados según sus quejas?
b) ¿Qué cantidad de clientes presentó 4 quejas?
c) ¿Qué por ciento de clientes presentó 2 quejas?
d) ¿Cuántos clientes expresaron hasta 5 quejas?
e) ¿Qué por ciento representan los clientes que presentaron hasta 3 quejas?

SOLUCIÓN

a) Fueron analizadas las quejas de 12 clientes
b) 3 clientes
c) El 8.3% de los clientes
d) 10 clientes
e) El 50%

Capítulo
3

Estadígrafos y gráficos.

3.1. Tipos de estadígrafos o estadísticos.

Ya se ha analizado el estudio de la distribución de frecuencias de un conjunto de datos, y cómo construir tablas para ordenar los mismos. No obstante, es posible hacer que este estudio sea más factible mediante el análisis cualitativo de un gráfico correspondiente, y la obtención de ciertas cantidades numéricas que permitan una mayor caracterización del conjunto de datos.

Estas cantidades numéricas son funciones de los valores que toma una variable de estudio, y en estadística, a este tipo general de función se le denomina estadígrafo o estadístico.

Existen 4 tipos de estadígrafos:
- los de posición
- los de dispersión
- los de forma
- los de concentración

3.2. Estadígrafos de posición.

Existen varios estadísticos que constituyen medidas de tendencia central de los datos. Dichas medidas son:
- media aritmética o promedio
- media armónica
- media geométrica
- mediana
- moda
- cuantiles

3.2.1. Media aritmética o promedio.

Para distribuciones de tipo II y III, se denomina media aritmética ponderada.

Ventajas de la media aritmética:
- es calculable en todas las variables siempre que las observaciones sean cuantitativas
- para su cálculo se utilizan todos los valores de la distribución
- es única para cada distribución de frecuencias
- al ser el centro de gravedad de la distribución, representa los valores observados

Desventajas de la media aritmética:
- es un valor muy sensible a los valores extremos de la distribución, por lo que en distribuciones de gran dispersión de datos, puede llegar a perder totalmente su significado

3.2.2. Media armónica.

Ventajas de la media armónica:
- es más representativa que otras medidas en los casos de obtener promedios de rendimientos, velocidades, productividades, tazas, etc.
- para su cálculo se tiene en cuenta todos los valores de la distribución
- los valores extremos tienen una menor influencia que en la media aritmética

Desventajas de la media armónica:
- cuando se utiliza para variables en las que hay valores muy pequeños, sus inversos pueden aumentar casi hasta el infinito eliminando el efecto del resto de los valores
- no es posible su cálculo cuando algún valor es cero, puesto que se produce una indeterminación matemática

3.2.3. Media geométrica.

Ventajas de la media geométrica:
- es más representativa que la media aritmética cuando la variable evoluciona de forma acumulativa con efectos multiplicatorios
- para su cálculo se tiene en cuenta todos los valores de la distribución
- los valores extremos tienen una menor influencia que en la media aritmética

Desventajas de la media geométrica:
- si hay observaciones nulas o negativas, no se puede calcular

3.2.4. Mediana.

Ventajas de la mediana:
- es la medida más representativa en el caso de las variables que sólo admiten una escala ordinal
- no es sensible a los valores extremos de la distribución a diferencia de las medias

Desventajas de la mediana:

- en su determinación no se tienen en cuenta todos los valores de la variable, aunque esto puede constituir una ventaja, ya que es posible su cálculo cuando no se conocen los valores extremos, pero sí su frecuencia

3.2.5. Moda.

En las distribuciones de tipo I, no tiene sentido hablar de moda, ya que las frecuencias absolutas (n_i) son todas unitarias.

Se dice que un valor de una variable constituye una moda relativa, cuando su frecuencia absoluta es mayor que la de los valores contiguos.
Puede, en una distribución, existir más de una moda.

Ventajas de la moda:

- es la única medida de posición central que puede obtenerse en las distribuciones con datos cualitativos, ya que es posible determinar la categoría o modalidad que más se repite en un determinado atributo

Desventajas de la moda:

- en su determinación no intervienen todos los valores de la distribución, centrándose sólo en la mayor frecuencia absoluta de un determinado valor de la variable

3.2.6. Cuantiles.

Son valores de la variable que dividen la distribución en partes iguales respecto a las frecuencias de la distribución.
Se obtienen cuartiles, deciles, percentiles, etc.

3.3. Estadígrafos de dispersión.

Existen varios estadísticos que constituyen medidas de dispersión o variabilidad de los datos. Dichas medidas son:

Absolutas:

- rango o recorrido
- desviación típica o estándar
- varianza
- recorrido intercuartílico
- desviación absoluta media

Relativas:

- coeficiente de variación de Pearson
- coeficiente de apertura
- recorrido relativo
- recorrido semintercuartílico

3.3.1. Desviación típica o estándar.

Una desviación estándar pequeña, significa que todos los valores de la distribución, se sitúan próximos a la media, mientras que una desviación típica elevada, significa la existencia de valores por exceso o por defecto, muy alejados de la media.

La tipificación de variables consiste en expresar la diferencia entre la media y los valores de la variable, en términos de desviación típica. Por ello, esta última medida de dispersión, es la más importante en Estadística.

3.3.2. Varianza.

Cuando la varianza está referida a una muestra y no al total de la población, se denomina cuasivarianza. Su valor es igual a la desviación típica elevada al cuadrado.

3.3.3. Coeficiente de variación de Pearson.

El coeficiente de variación representa el número de veces que la desviación típica contiene a la media.

Es una medida de dispersión relativa que permite hacer comparaciones entre distribuciones diferentes, o sea, que no vengan expresadas en la misma unidad de medida.

Valores menores de la unidad, indican que el promedio representa adecuadamente a la distribución de frecuencias, ya que la dispersión es inferior a la media aritmética. A partir de la unidad, hay que rechazar la media aritmética como parámetro representativo de los datos de la distribución.

3.4. Estadígrafos de forma.

Existen varios estadísticos que constituyen medidas de forma de los datos. Dichas medidas son:

- coeficiente de asimetría
- coeficiente de curtosis o apuntamiento

3.4.1. Asimetría.

La simetría es importante para saber si los valores de la variable, se concentran en una determinada zona del recorrido de la misma. Existe la distribución asimétrica negativa a la izquierda, la distribución asimétrica positiva a la derecha, y la distribución simétrica.

3.4.2. Curtosis o apuntamiento.

Para estudiar el apuntamiento o curtosis, se toma la distribución normal como referencia. Cuando una distribución es más apuntada que la normal, se dice que es leptocúrtica, y si es menos apuntada, entonces es platicúrtica. Por último, si posee igual apuntamiento que la normal, entonces se denomina mesocúrtica.

3.5. Estadígrafos de concentración.

Existen varios estadísticos que constituyen medidas de concentración de los datos. Dichas medidas son:

- índice de Gini
- curva de Lorenz

Miden el mayor o menor grado de equidad en el reparto de las variables.

Obsérvese un ejemplo.

Ejemplo 1:

Un grupo de 20 estudiantes de la Licenciatura en Turismo, fue sometido al examen final de la asignatura Marketing Turístico la semana pasada. Las notas oscilaron entre 2 puntos (desaprobado) y 5 puntos (excelente). Los datos se muestran a continuación:

5	5	4	2	3
4	4	4	5	2
2	4	3	3	5
4	4	5	5	3

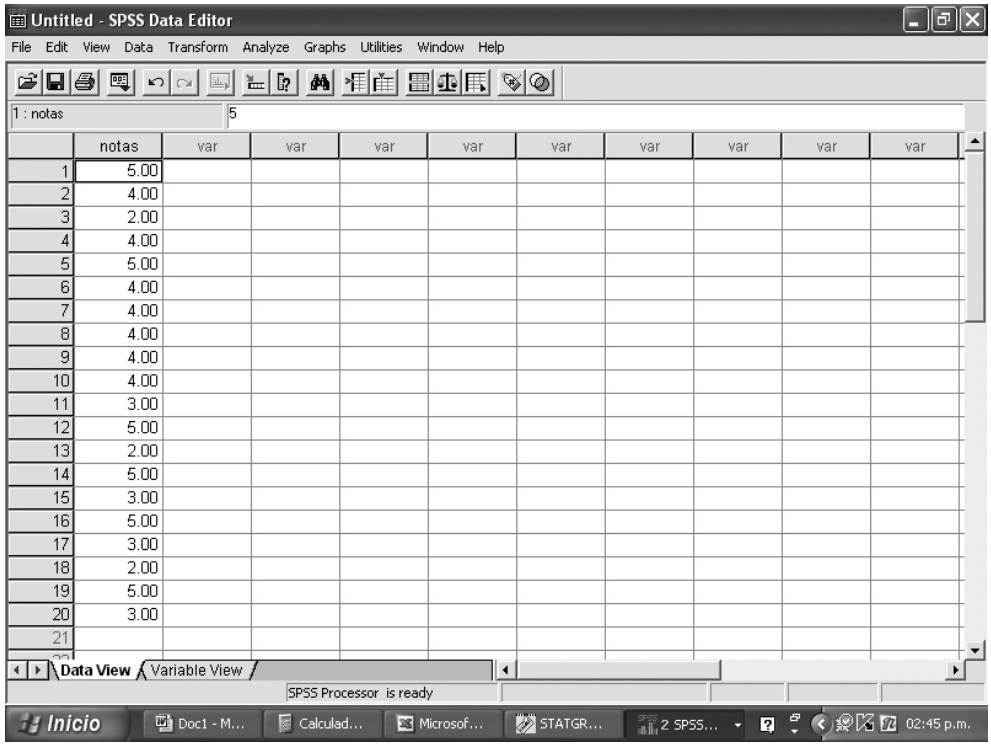
Se desea conocer con dichos datos de la variable:

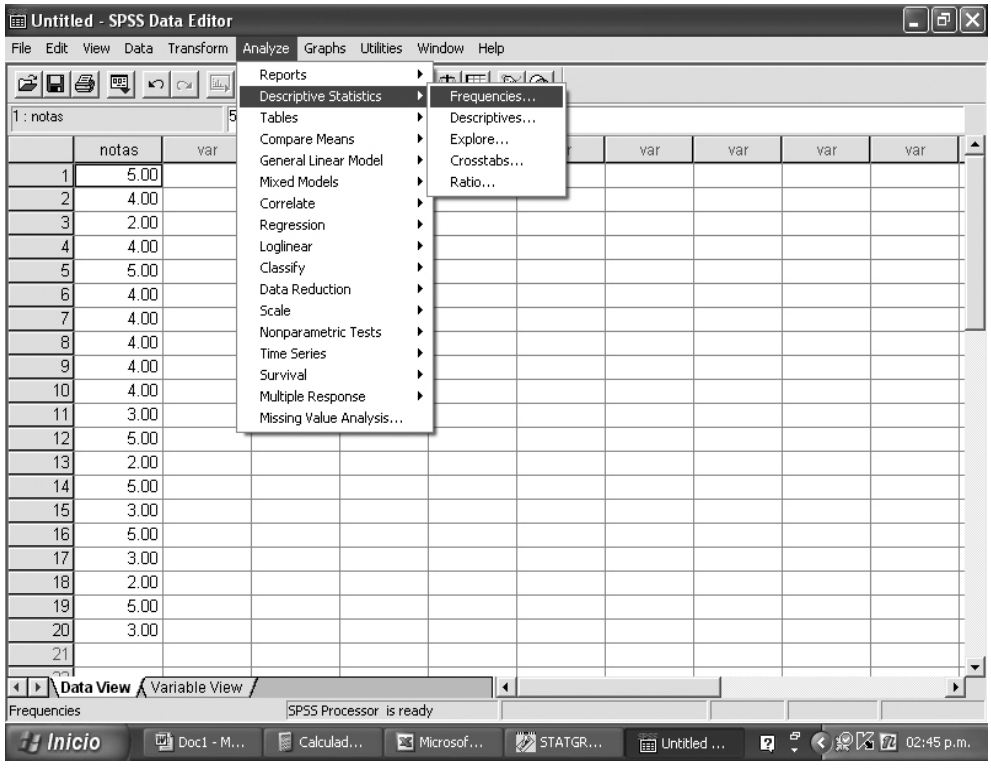
a) La media aritmética o promedio
b) La mediana
c) La moda
d) El rango de la variable
e) La desviación estándar
f) La varianza
g) La asimetría
h) La curtosis

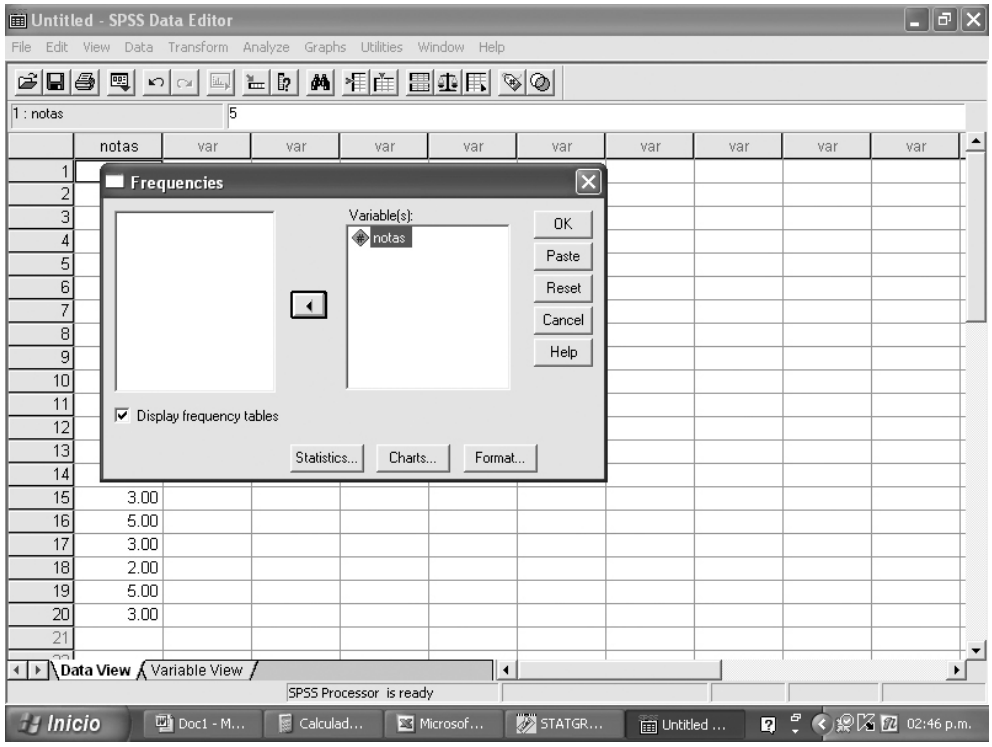
Solución:

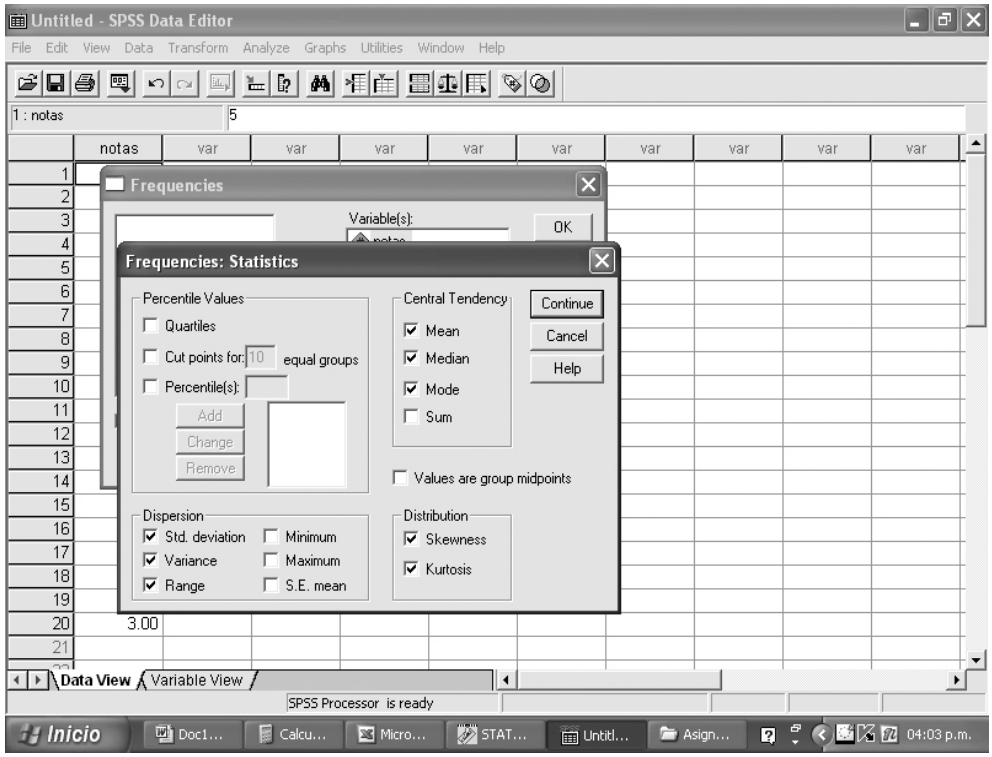
Variable cuantitativa: notas.

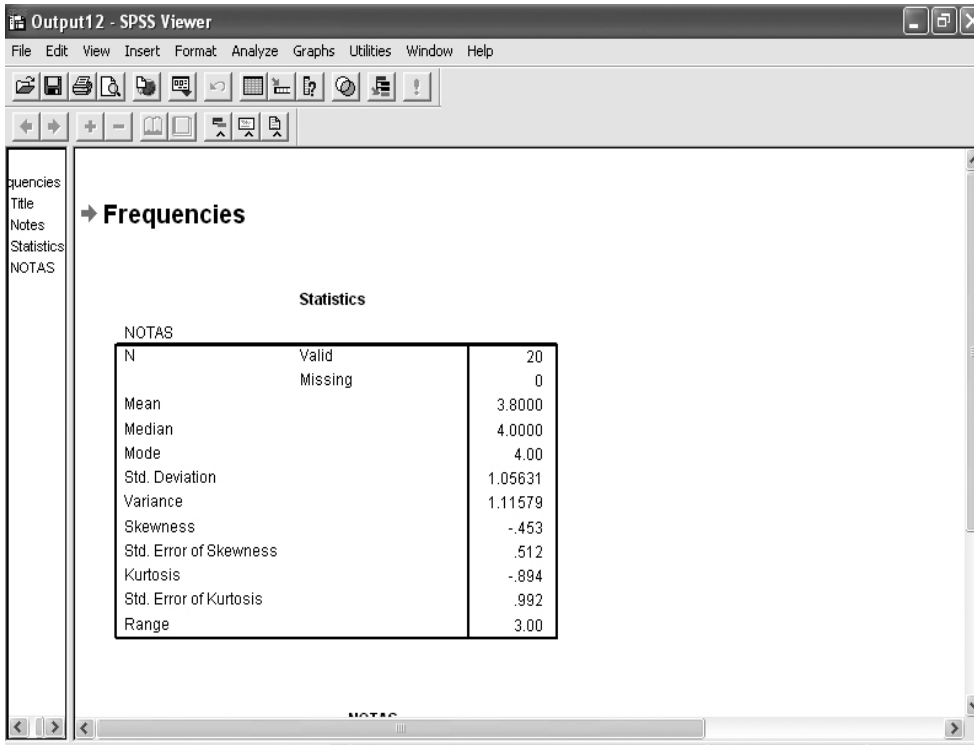
Empleando el SPSS, sería:


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : notas
5
notas
var
Data View
Variable View
SPSS Processor is ready
Inicio
Doc1 - M...
Calculad...
Microsof...
STATGR...
2 SPSS...
02:45 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Frequencies...
Descriptives...
Explore...
Crosstabs...
Ratio...
Data View
Variable View
Frequencies
SPSS Processor is ready
Inicio
Doc1 - M...
Calculad...
Microsof...
STATGR...
Untitled ...
02:45 p.m.

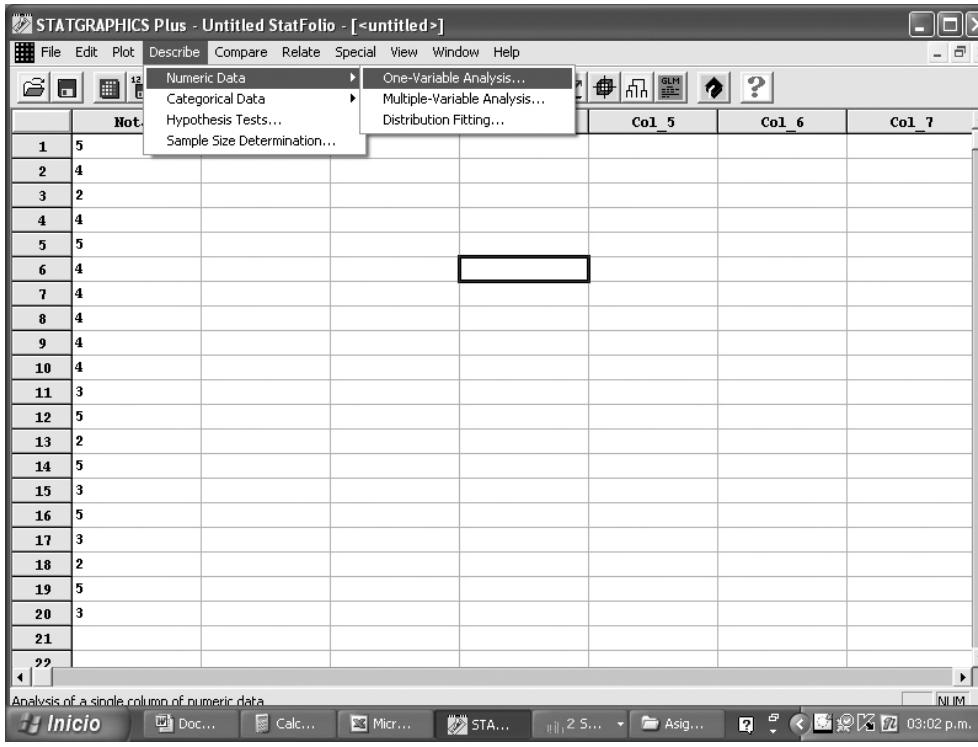

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : notas
5
notas
var
Frequencies
Variable(s):
notas
OK
Paste
Reset
Cancel
Help
Display frequency tables
Statistics...
Charts...
Format...
3.00
5.00
3.00
2.00
5.00
3.00
Data View
Variable View
SPSS Processor is ready
Inicio
Doc1 - M...
Calculad...
Microsof...
STATGR...
Untitled ...
02:46 p.m.

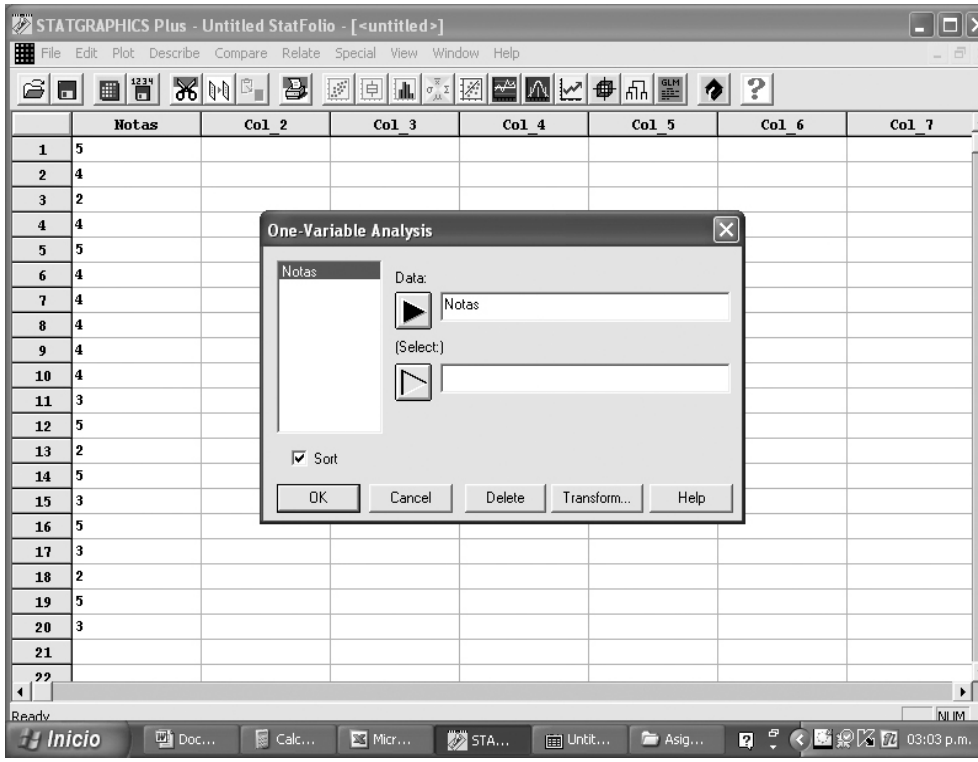

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : notas
5
notas
var
Frequencies
Variable(s):
OK
Frequencies: Statistics
Percentile Values
Quartiles
Cut points for: 10 equal groups
Percentile(s):
Add
Change
Remove
Central Tendency
Mean
Median
Mode
Sum
Continue
Cancel
Help
Values are group midpoints
Dispersion
Std. deviation
Minimum
Variance
Maximum
Range
S.E. mean
Distribution
Skewness
Kurtosis
3.00
Data View
Variable View
SPSS Processor is ready
Inicio
Doc1...
Calcu...
Micro...
STAT...
Untitl...
Asign...
04:03 p.m.

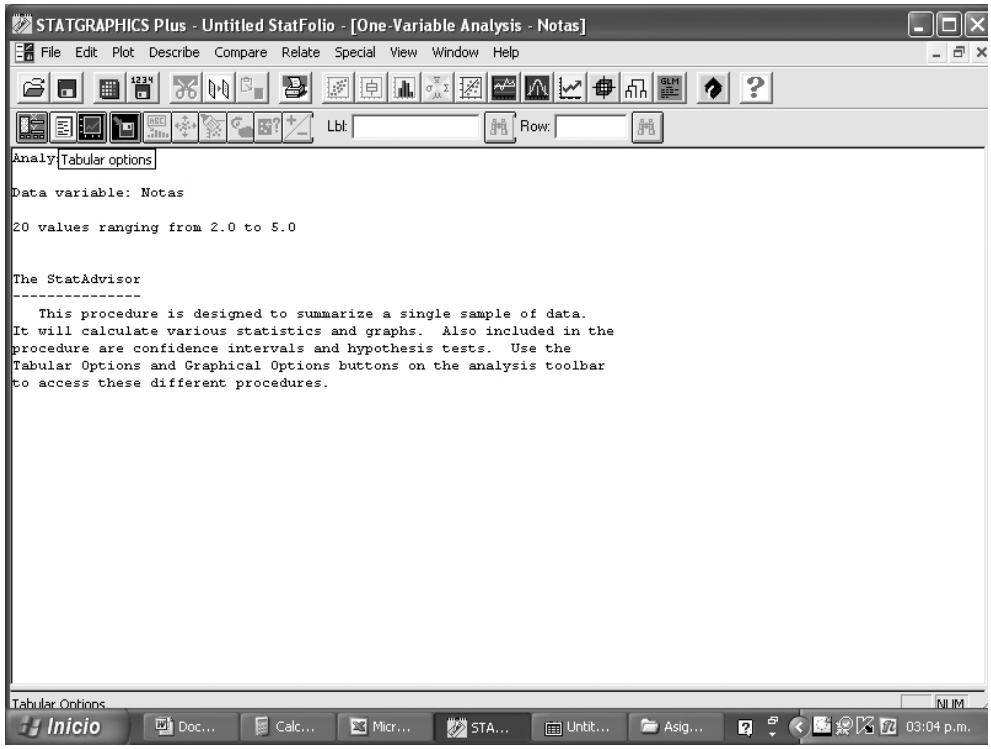


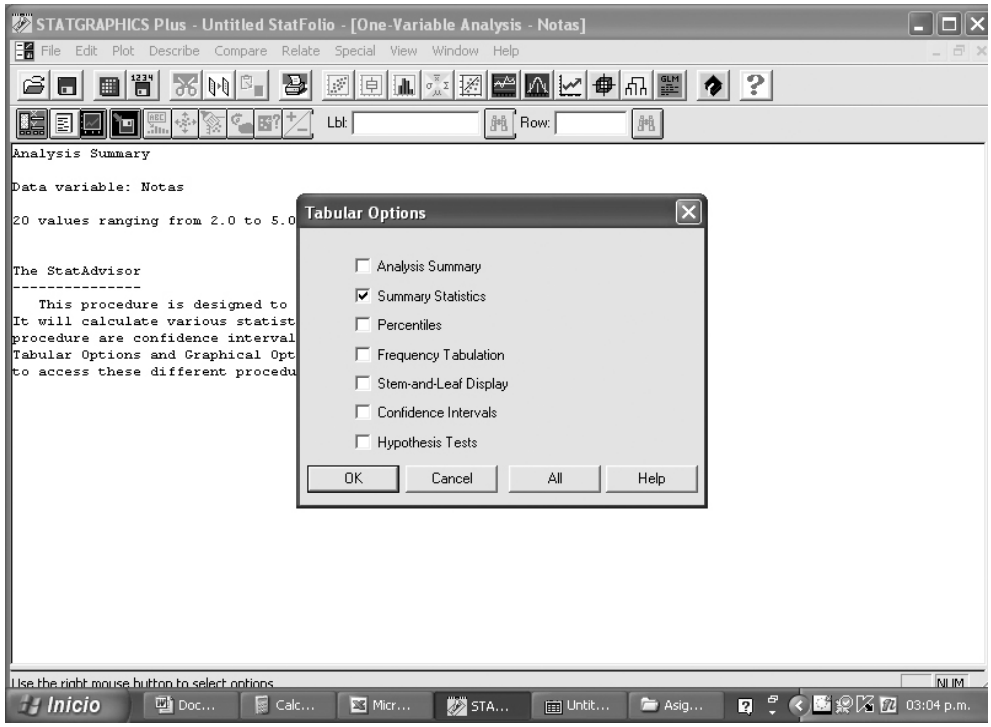
a) La media aritmética de las notas es 3.8 (de regular a bien)
b) La mediana de las notas es 4 (bien)
c) La moda de las notas es 4 (bien)
d) El rango es de 3 pues los datos de las notas oscilan entre 2 y 5
e) La desviación típica equivale a 1.06
f) La varianza es de 1.12
g) La asimetría es igual a -0.45 por lo cual la distribución de datos es asimétrica a la izquierda
h) La curtosis es igual a -0.89 por lo cual la distribución es platicúrtica

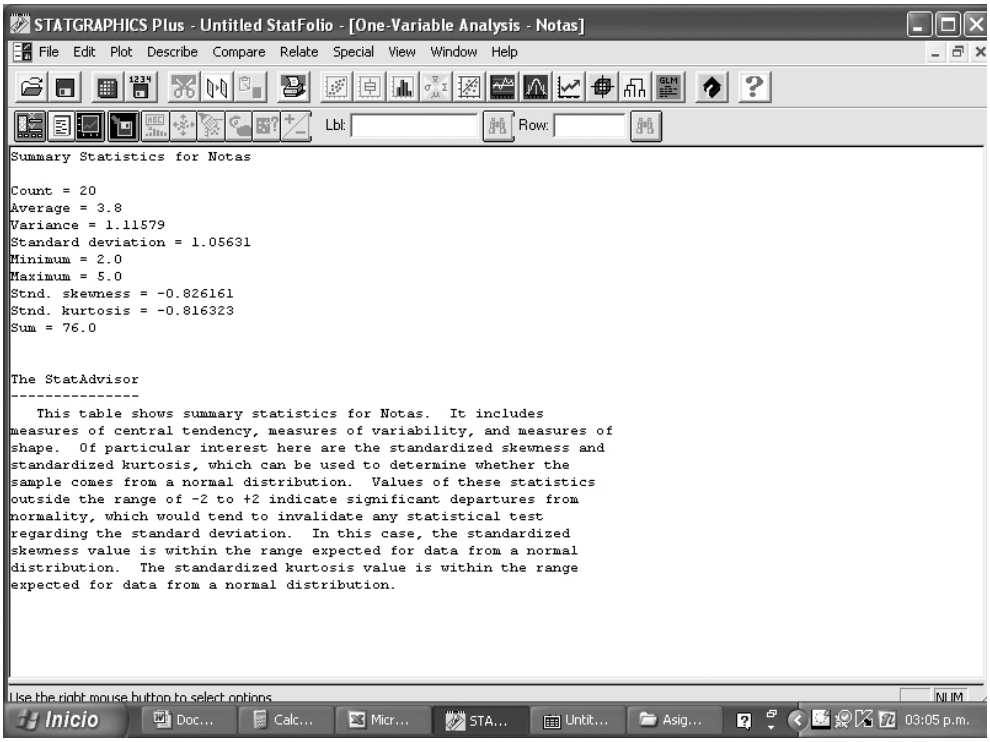
Para obtener algunas de las medidas de tendencia central, dispersión y forma utilizando el STATGRAPHICS Plus, se opera de la siguiente manera:


STATGRAPHICS Plus - Untitled StatFolio - [<untitled>]
File Edit Plot Describe Compare Relate Special View Window Help
Numeric Data
Categorical Data
Hypothesis Tests...
Sample Size Determination...
One-Variable Analysis...
Multiple-Variable Analysis...
Distribution Fitting...
Col_5 Col_6 Col_7
Analysis of a single column of numeric data
NUM
Inicio
03:02 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [<untitled>]
Notas Col_2 Col_3 Col_4 Col_5 Col_6 Col_7
One-Variable Analysis
Notas
Data:
Notas
(Select:)
Sort
OK
Cancel
Delete
Transform...
Help
Ready
NUM
Inicio
03:03 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [One-Variable Analysis - Notas]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Analy Tabular options
Data variable: Notas
20 values ranging from 2.0 to 5.0
The StatAdvisor
This procedure is designed to summarize a single sample of data.
It will calculate various statistics and graphs. Also included in the
procedure are confidence intervals and hypothesis tests. Use the
Tabular Options and Graphical Options buttons on the analysis toolbar
to access these different procedures.
Tabular Options
NUM
Inicio
Doc...
Calc...
Micr...
STA...
Untit...
Asig...
03:04 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [One-Variable Analysis - Notas]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Analysis Summary
Data variable: Notas
20 values ranging from 2.0 to 5.0
The StatAdvisor
This procedure is designed to
It will calculate various statist
procedure are confidence interval
Tabular Options and Graphical Opt
to access these different procedu
Tabular Options
Analysis Summary
Summary Statistics
Percentiles
Frequency Tabulation
Stem-and-Leaf Display
Confidence Intervals
Hypothesis Tests
OK
Cancel
All
Help
Use the right mouse button to select options
NUM
Inicio
Doc...
Calc...
Micr...
STA...
Untit...
Asig...
03:04 p.m.



3.6. Gráficos.

Para hacer más comprensible la información estadística, es muy útil representar las distribuciones gráficamente.

Existen varios tipos de gráficos:
- diagrama de Pareto
- gráfico de barras
- histograma
- gráfico de series temporales
- gráfico de sectores

3.6.1. Diagrama de Pareto.

Este gráfico:
- se emplea para representar datos cualitativos
- se ordenan las clases o categorías según la frecuencia relativa (f_i) de su aparición
- cada clase se representa por un rectángulo con una altura igual a la frecuencia relativa

3.6.2. Gráfico de barras.

Este gráfico:
- se emplea para variables discretas en distribuciones de frecuencias de datos sin agrupar

3.6.3. Histograma.

Este gráfico:
- es la representación más frecuente para ver los datos agrupados
- es un conjunto de rectángulos donde cada uno representa una clase
- la base de los rectángulos es igual a la amplitud del intervalo, y la altura, representa la frecuencia de cada clase

3.6.4. Gráfico de series temporales.

En este gráfico:
- se representan los valores ordenados según la secuencia temporal

3.6.5. Gráfico de sectores.

Este gráfico:
- se utiliza para mostrar las contribuciones relativas de cada punto de los datos al total de la serie
- sólo se representa una serie

Véase un ejemplo.

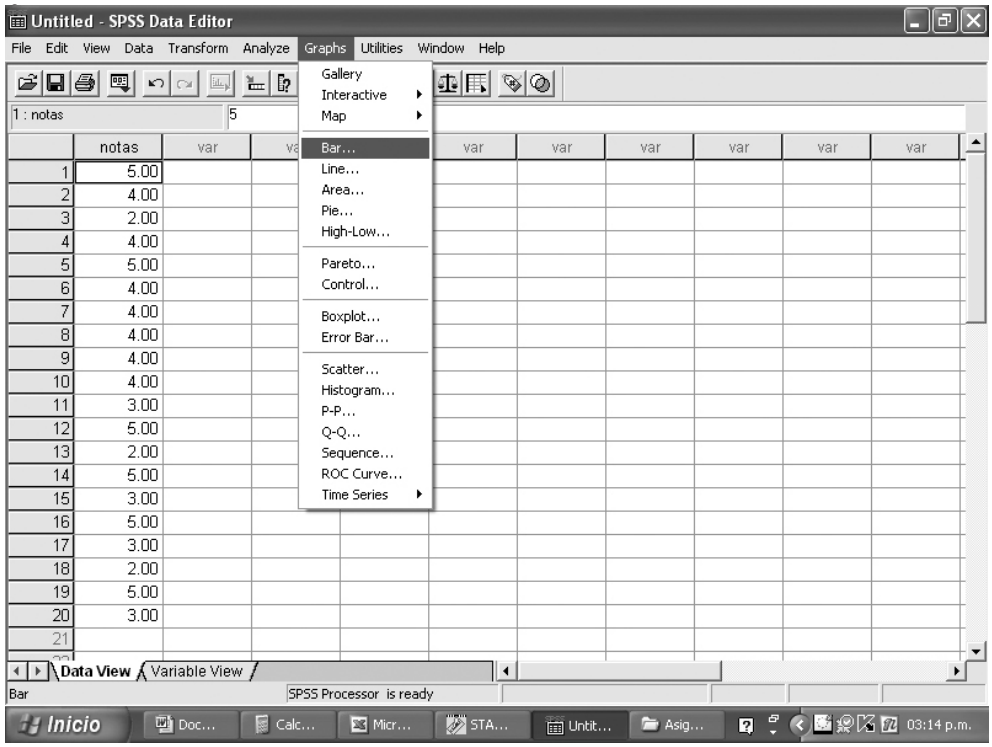
Ejemplo 2:

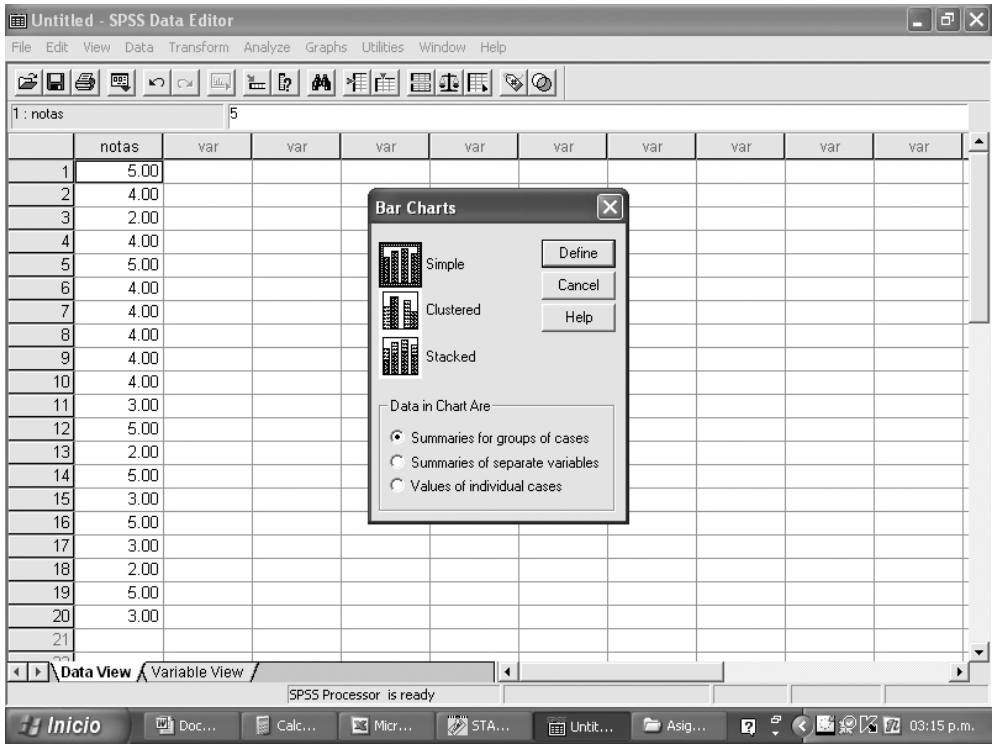
Para obtener una información gráfica del comportamiento de los datos de la variable "notas", se puede elaborar:

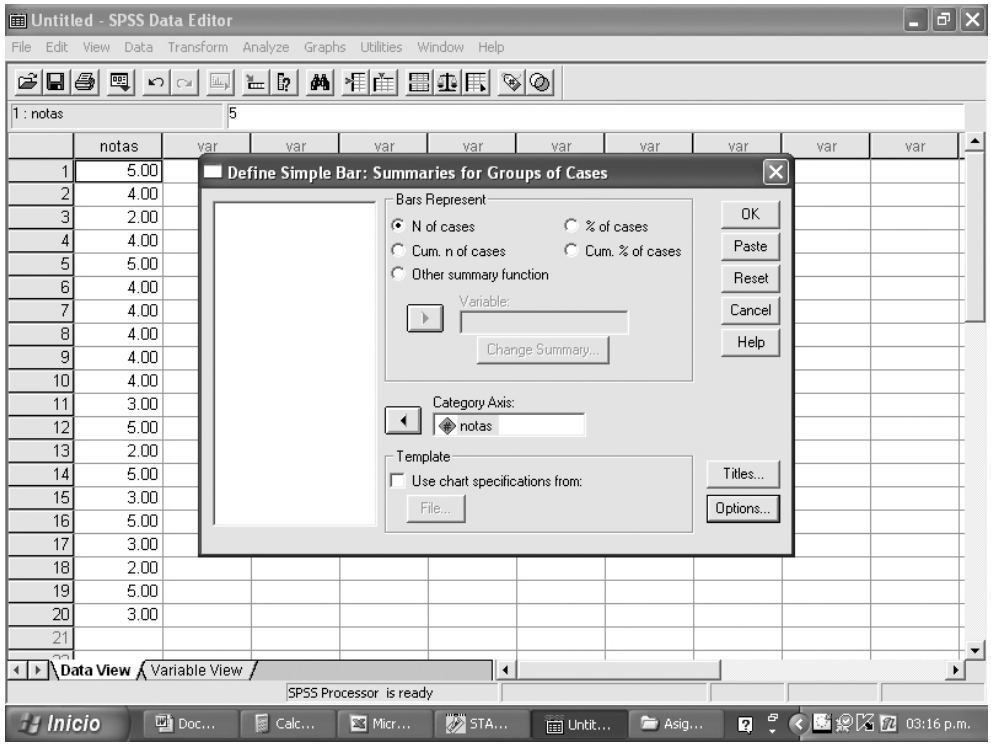
a) un gráfico de barras
b) un histograma de frecuencias con la curva normal dibujada

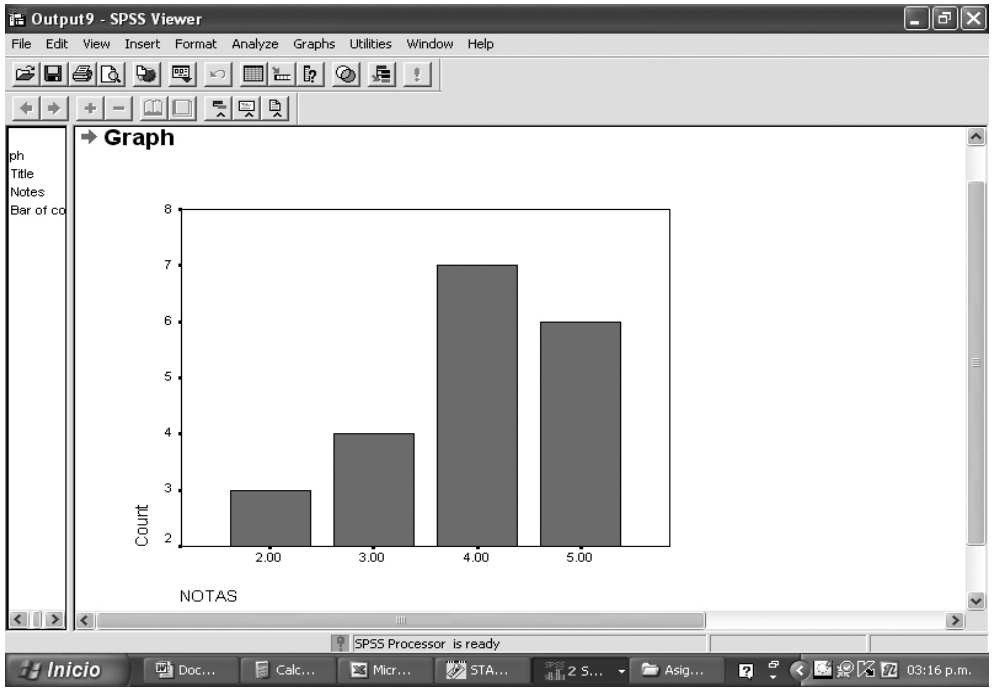
Solución:

Para elaborar un gráfico de barras empleando el SPSS, se procede de la siguiente


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Gallery
Interactive
Map
Bar...
Line...
Area...
Pie...
High-Low...
Pareto...
Control...
Boxplot...
Error Bar...
Scatter...
Histogram...
P-P...
Q-Q...
Sequence...
ROC Curve...
Time Series
1 : notas
5
notas
var
Data View
Variable View
Bar
SPSS Processor is ready
Inicio
Doc...
Calc...
Micr...
STA...
Untit...
Asig...
03:14 p.m.

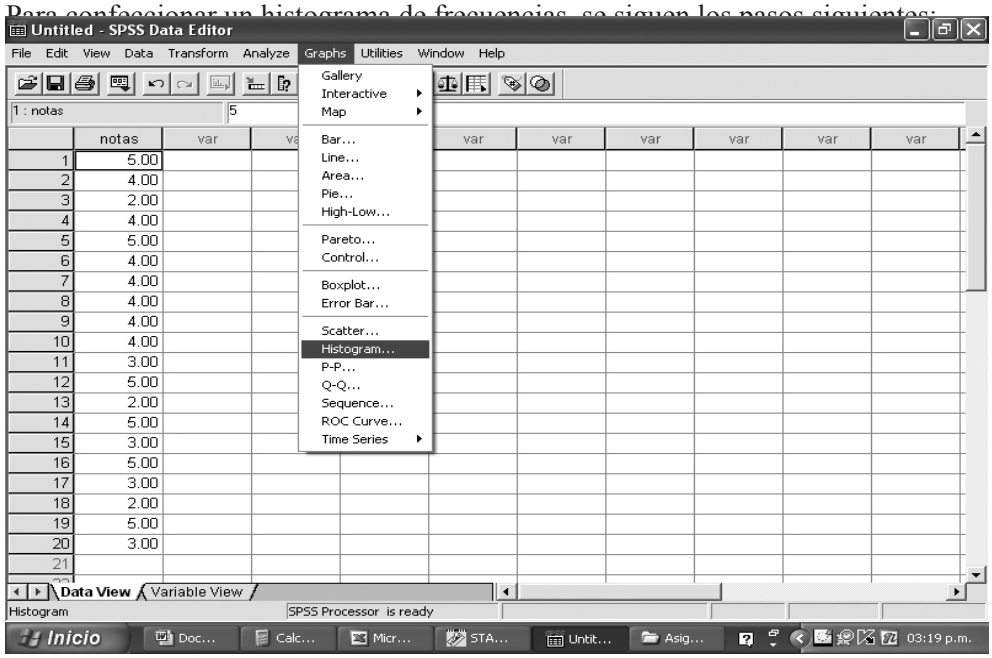

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : notas
5
notas
var
Bar Charts
Simple
Clustered
Stacked
Define
Cancel
Help
Data in Chart Are
Summaries for groups of cases
Summaries of separate variables
Values of individual cases
Data View
Variable View
SPSS Processor is ready
Inicio
03:15 p.m.

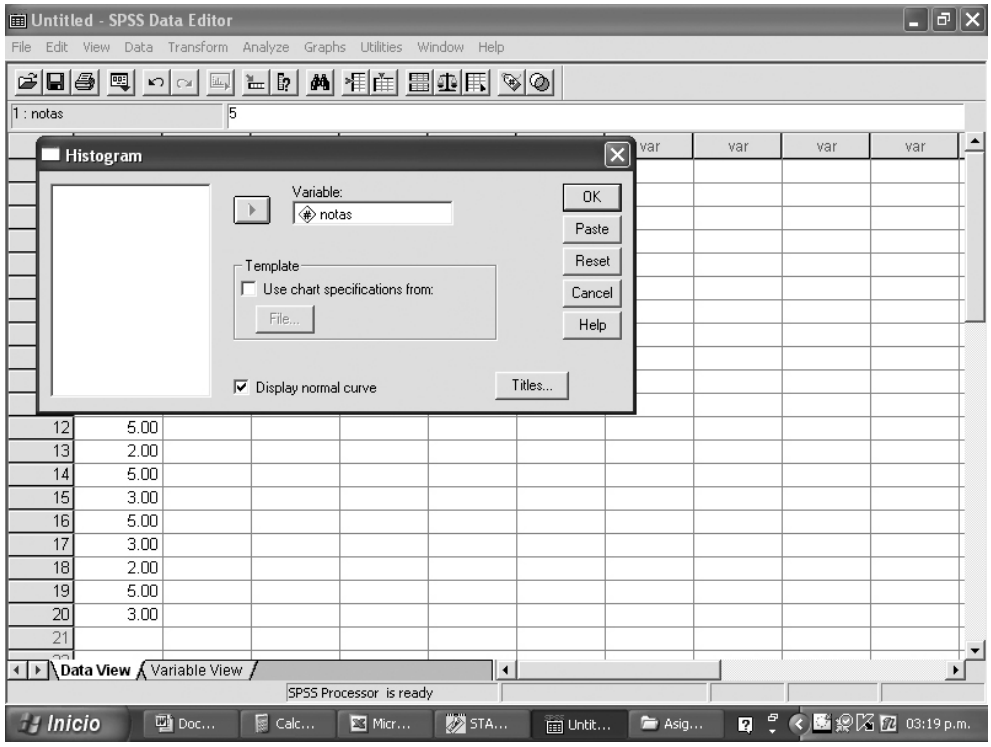

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : notas
5
notas
var
Define Simple Bar: Summaries for Groups of Cases
Bars Represent
N of cases
% of cases
Cum. n of cases
Cum. % of cases
Other summary function
Variable:
Change Summary...
OK
Paste
Reset
Cancel
Help
Category Axis:
notas
Template
Use chart specifications from:
File...
Titles...
Options...
Data View
Variable View
SPSS Processor is ready
Inicio
03:16 p.m.

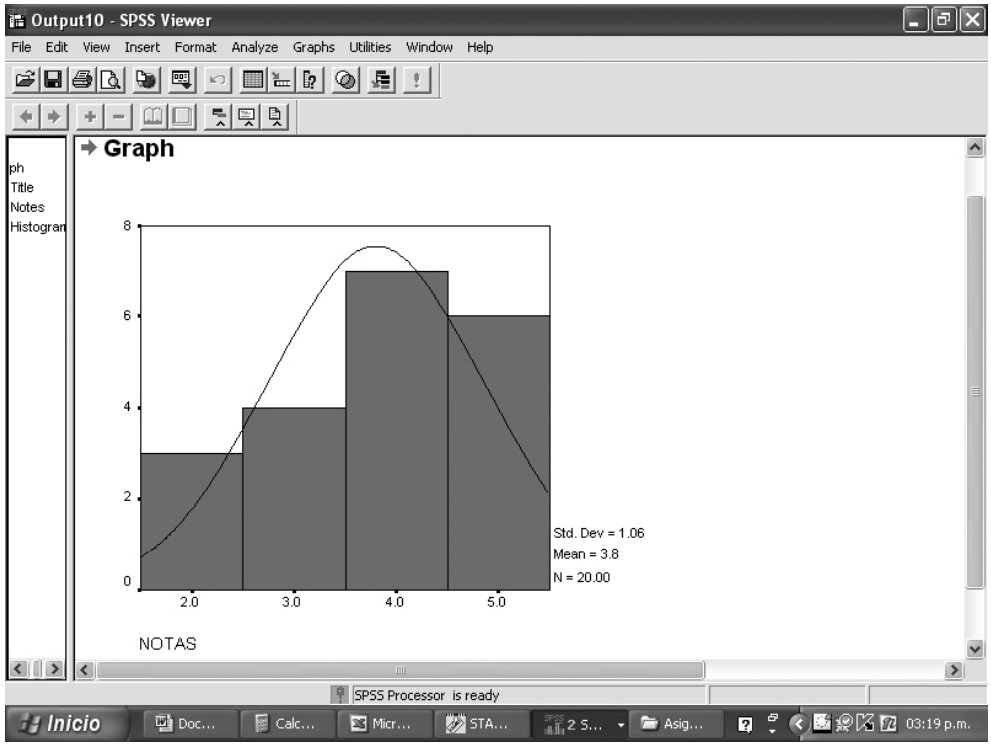


a) Véase que tres estudiantes desaprobaron el examen (2 puntos), cuatro estudiantes recibieron puntuación de 3 (regular), siete recibieron la calificación de 4 puntos (bien) y finalmente, seis estudiantes obtuvieron resultados excelentes (5 puntos).

Para confeccionar un histograma de frecuencias, se siguen los pasos siguientes:

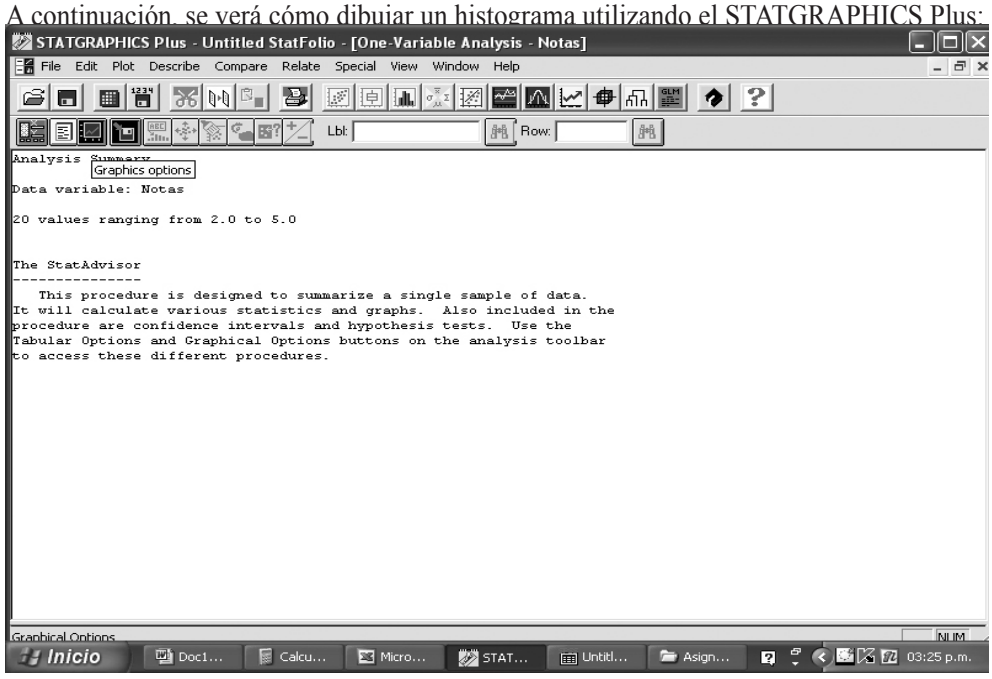


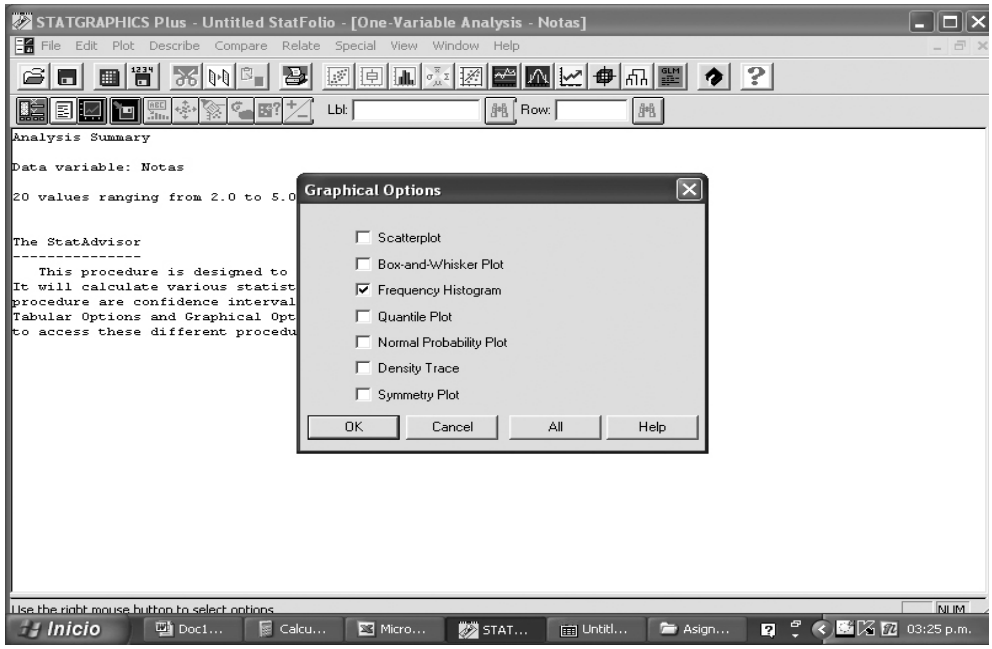

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : notas
5
Histogram
Variable:
notas
OK
Paste
Reset
Cancel
Help
Template
Use chart specifications from:
File...
Display normal curve
Titles...
var
var
var
var
12 5.00
13 2.00
14 5.00
15 3.00
16 5.00
17 3.00
18 2.00
19 5.00
20 3.00
21
Data View
Variable View
SPSS Processor is ready
Inicio
Doc...
Calc...
Micr...
STA...
Untit...
Asig...
03:19 p.m.

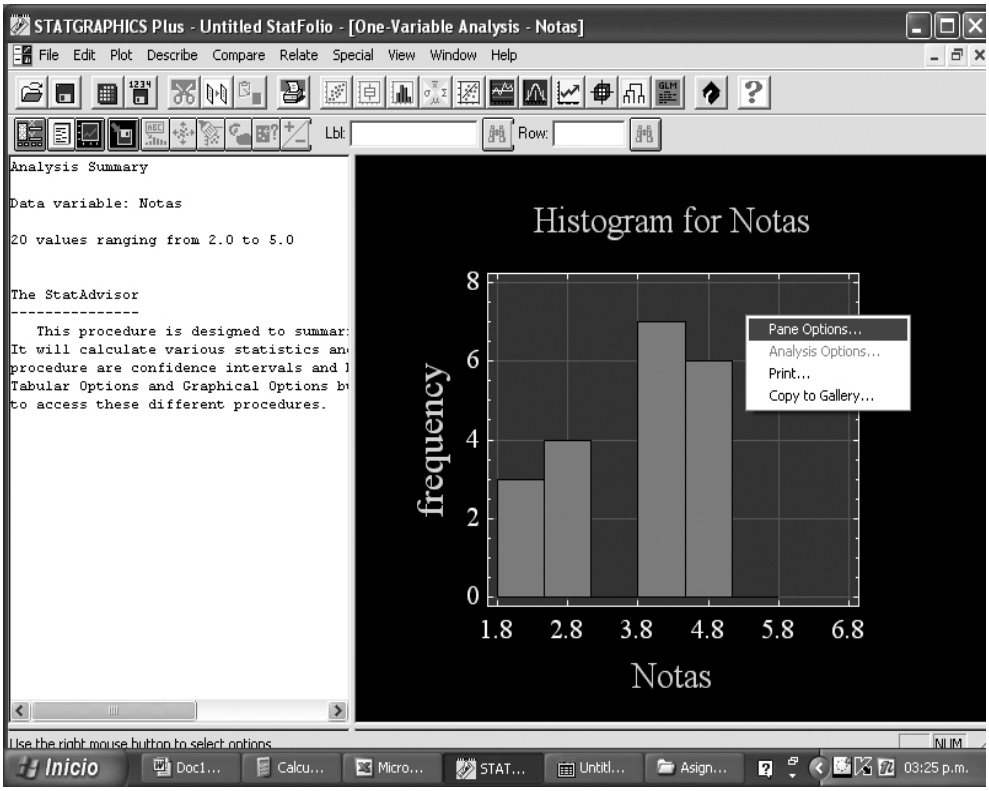

Output10 - SPSS Viewer
File Edit View Insert Format Analyze Graphs Utilities Window Help
ph
Title
Notes
Histogram
Graph
8
6
4
2
0
2.0
3.0
4.0
5.0
NOTAS
Std. Dev = 1.06
Mean = 3.8
N = 20.00
SPSS Processor is ready
Inicio
Doc...
Calc...
Micr...
STA...
2 S...
Asig...
03:19 p.m.

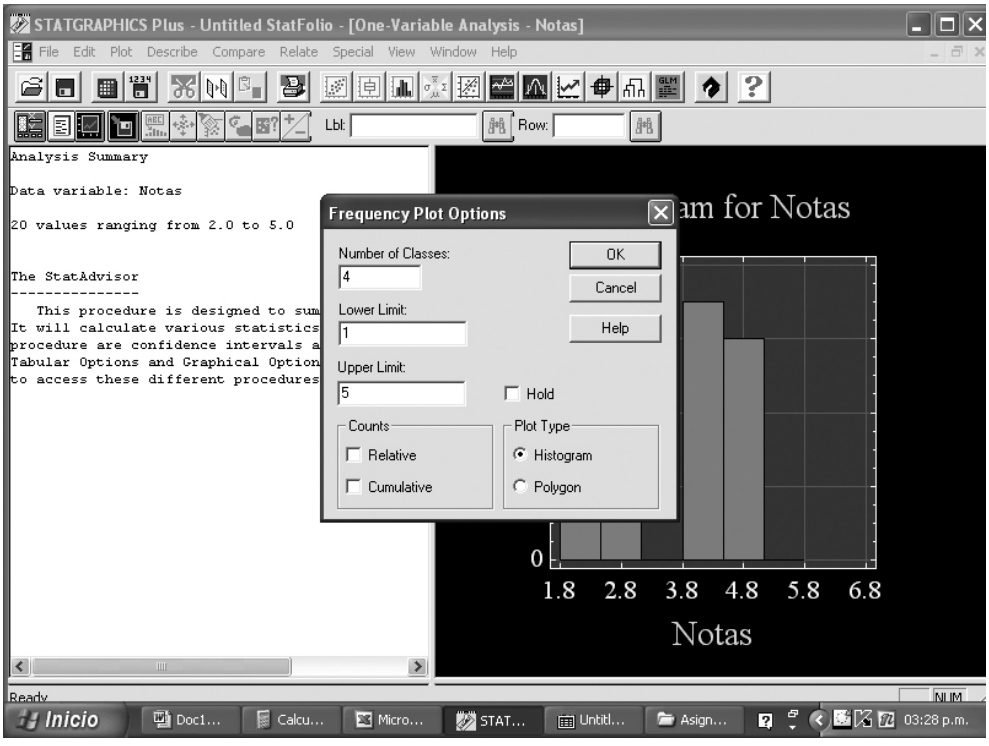
b) Obsérvese que un histograma es muy similar a un gráfico de barras, pero comúnmente, permite añadir una curva que representa la distribución normal, la cual admite conocer, de forma visual, el comportamiento de las medidas de tendencia central de los datos de la variable, de dispersión, y específicamente, las medidas de forma (asimetría y curtosis).

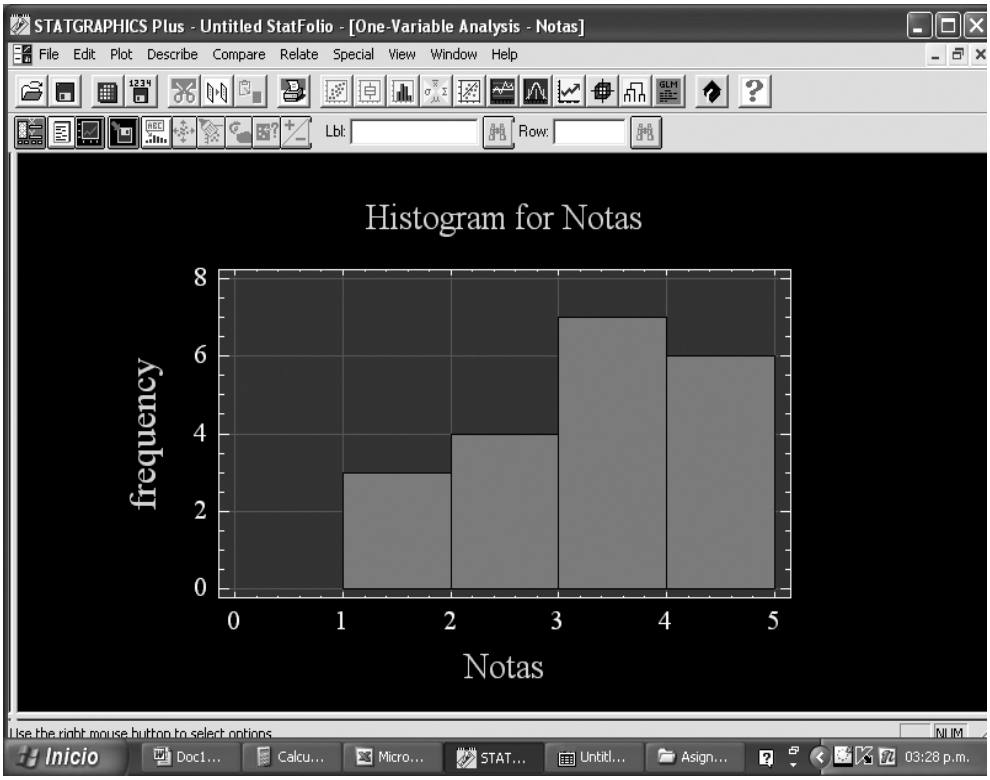
A continuación, se verá cómo dibujar un histograma utilizando el STATGRAPHICS Plus:






STATGRAPHICS Plus - Untitled StatFolio - [One-Variable Analysis - Notas]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Analysis Summary
Data variable: Notas
20 values ranging from 2.0 to 5.0
The StatAdvisor
This procedure is designed to summar
It will calculate various statistics an
procedure are confidence intervals and
Tabular Options and Graphical Options b
to access these different procedures.
Histogram for Notas
frequency
Notas
Pane Options...
Analysis Options...
Print...
Copy to Gallery...
Use the right mouse button to select options
Inicio
03:25 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [One-Variable Analysis - Notas]
Analysis Summary
Data variable: Notas
20 values ranging from 2.0 to 5.0
The StatAdvisor
Frequency Plot Options
Number of Classes:
4
Lower Limit:
1
Upper Limit:
5
OK
Cancel
Help
Hold
Counts
Relative
Cumulative
Plot Type
Histogram
Polygon
Notas
Ready
Inicio
03:28 p.m.



Obsérvese que:

- tres estudiantes desaprobaron
- cuatro estudiantes recibieron calificación de 3 puntos (regular)
- siete obtuvieron 4 puntos (bien)
- seis estudiantes obtuvieron calificación excelente (5 puntos)

EJERCITACIÓN

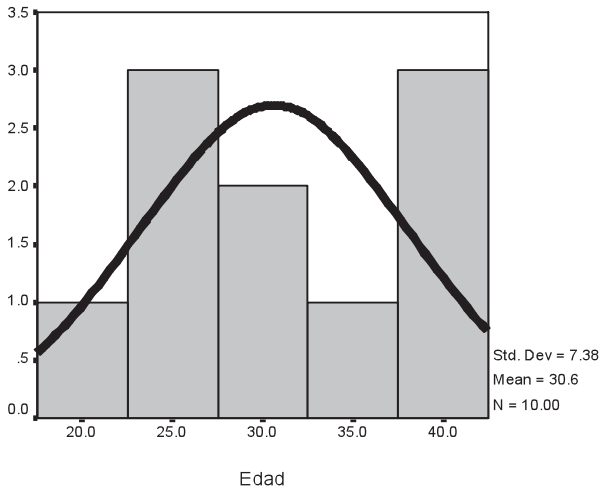
En la Agencia de Viajes Y, se ha recopilado la edad de un grupo de clientes que compraron la excursión "Crucero del Sol" la semana pasada. Los datos son los siguientes:

19	41	33	40	29
25	26	38	31	24

a) ¿Cuál es el promedio de edad del grupo de clientes?
b) ¿Cuál es el valor de la desviación típica de la edad del grupo?
c) ¿La distribución de frecuencias de los datos de edad, sigue una distribución normal según su curtosis?

SOLUCIÓN

a) La edad promedio del grupo es de 30.6 años
b) La desviación estándar de la edad equivale a 7.4
c) El coeficiente de apuntamiento es igual a -1.11 por lo cual la distribución de frecuencias de los datos de edad, es platicúrtica, o sea, no sigue una distribución normal



Capítulo
4

Pruebas de hipótesis paramétricas.

4.1. Generalidades acerca de las dócimas de hipótesis paramétricas.

Las dócimas de hipótesis paramétricas, hacen inferencia acerca del comportamiento de un parámetro medido en una población, por medio de estimar estadígrafos en una muestra. Los parámetros son la media, la varianza y la proporción de los datos de una variable.

El primer paso consiste en enunciar la hipótesis nula y la alternativa. En segundo lugar, calcular el estadígrafo de prueba. Como tercer paso, plantear la región crítica. El cuarto consiste en tomar la decisión estadística, y por último, dar la respuesta práctica según el caso de estudio que se trate.

Al utilizar softwares estadísticos para realizar una dócima de hipótesis, el investigador tiene que plantear las hipótesis o sospechas teóricas, y declarar con qué nivel de confiabilidad llevará a cabo la prueba. El resto, lo realiza el programa estadístico.

4.2. Dócima de hipótesis de la media.

Obsérvese un ejemplo.

Ejemplo 1:

Un grupo de estudiantes de segundo año de Licenciatura en Turismo, realizó un estudio acerca de los tiempos de duración de varios sub-procesos del servicio de Alimentos y Bebidas de un hotel, ubicado en el polo turístico de Varadero, y la relación existente de dichos tiempos con la satisfacción de los clientes.

Históricamente, los sub-procesos de Alimentos y Bebidas en ese hotel, han mantenido una media de duración de 2 minutos como máximo, lo cual ha

garantizado elevados índices de satisfacción de los clientes, puesto que éstos no han percibido demoras en el servicio, sin embargo, los estudiantes sospechan que últimamente esa duración promedio está superando los 2 minutos.

Para verificar si los tiempos de duración de los sub-procesos de Alimentos y Bebidas se han comportado como siempre, los estudiantes de Turismo decidieron tomar una muestra de 38 días, y en cada uno, realizaron una observación minuciosa determinando que el tiempo promedio que demoraba la ocurrencia de esos sub-procesos, era de 3 minutos con una varianza de 0,5 minutos^2.
Se desea determinar con un nivel de confiabilidad del 99%, si se verifica o no la sospecha de los estudiantes.

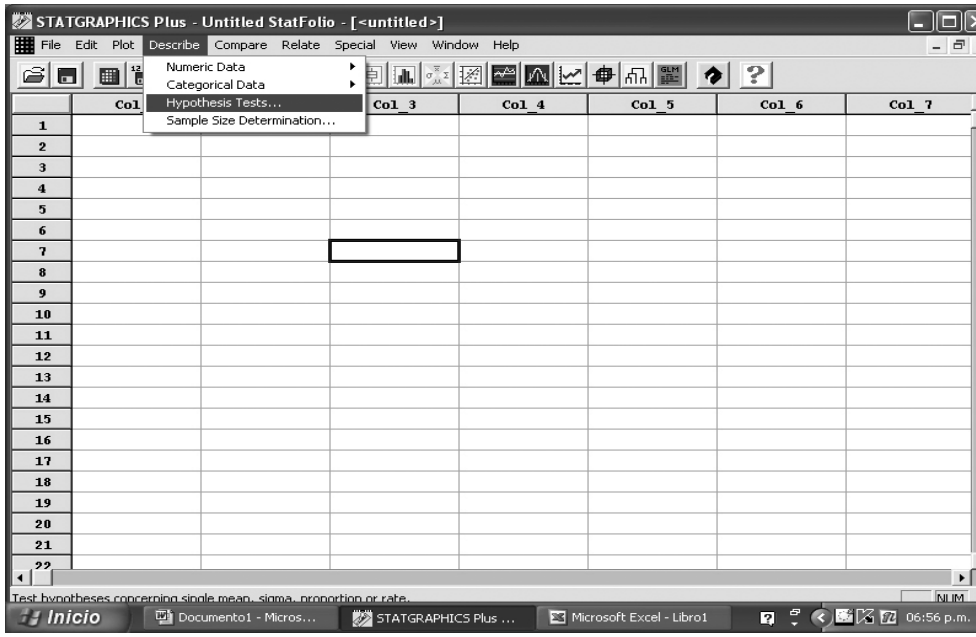
Solución:

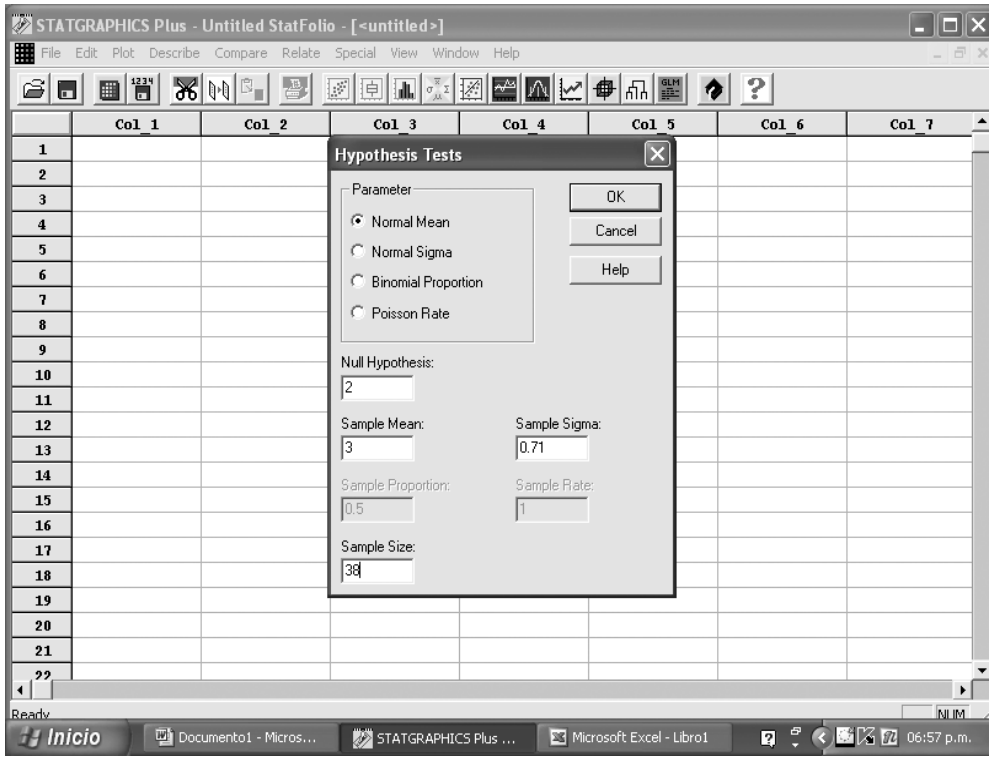
Véase que este es un ejemplo donde se realizará una dócima de hipótesis acerca de la **media o promedio** de la variable "tiempo de duración de los sub-procesos de Alimentos y Bebidas".

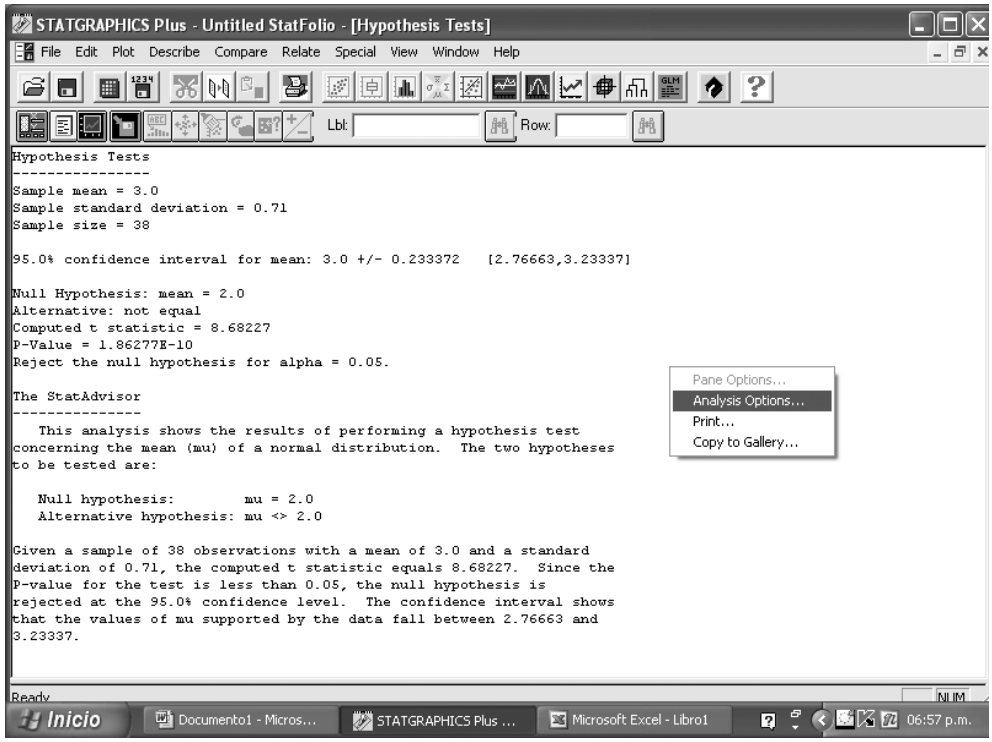
H_0: $\mu \leq 2$ (promedio menor o igual a 2 minutos)

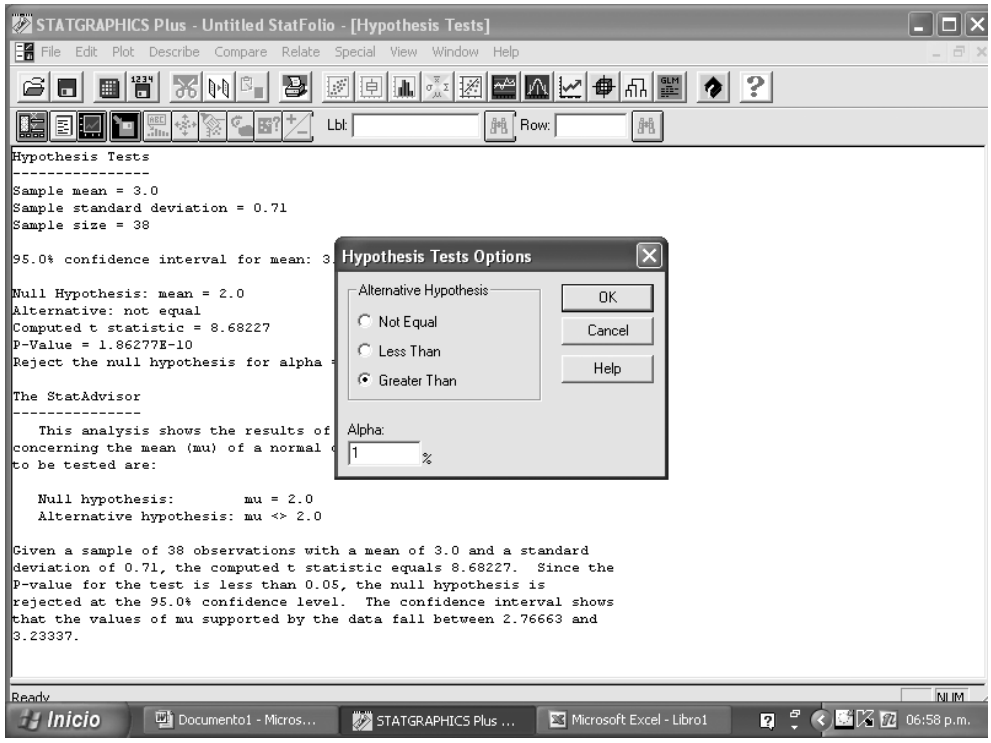
H_1: $\mu > 2$ (promedio superior a 2 minutos)

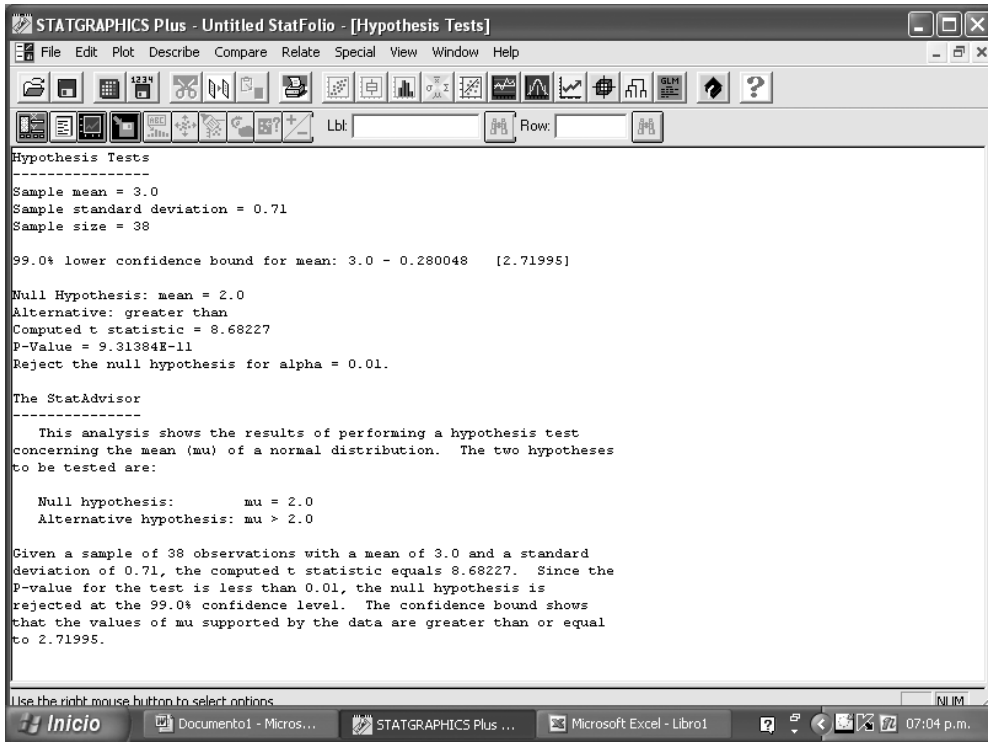
Utilizando el STATGRAPHICS, sería:




STATGRAPHICS Plus - Untitled StatFolio - [<untitled>]
File Edit Plot Describe Compare Relate Special View Window Help
Col_1 Col_2 Col_3 Col_4 Col_5 Col_6 Col_7
Hypothesis Tests
Parameter
Normal Mean
Normal Sigma
Binomial Proportion
Poisson Rate
OK
Cancel
Help
Null Hypothesis:
2
Sample Mean:
3
Sample Sigma:
0.71
Sample Proportion:
0.5
Sample Rate:
1
Sample Size:
38
Ready
NUM
Inicio
Documento1 - Micros...
STATGRAPHICS Plus ...
Microsoft Excel - Libro1
06:57 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [Hypothesis Tests]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl:
Row:
Hypothesis Tests
Sample mean = 3.0
Sample standard deviation = 0.71
Sample size = 38
95.0% confidence interval for mean: 3.0 +/- 0.233372 [2.76663,3.23337]
Null Hypothesis: mean = 2.0
Alternative: not equal
Computed t statistic = 8.68227
P-Value = 1.86277E-10
Reject the null hypothesis for alpha = 0.05.
The StatAdvisor
This analysis shows the results of performing a hypothesis test concerning the mean (mu) of a normal distribution. The two hypotheses to be tested are:
Null hypothesis: mu = 2.0
Alternative hypothesis: mu <> 2.0
Given a sample of 38 observations with a mean of 3.0 and a standard deviation of 0.71, the computed t statistic equals 8.68227. Since the P-value for the test is less than 0.05, the null hypothesis is rejected at the 95.0% confidence level. The confidence interval shows that the values of mu supported by the data fall between 2.76663 and 3.23337.
Pane Options...
Analysis Options...
Print...
Copy to Gallery...
Ready
NUM
Inicio
Documento1 - Micros...
STATGRAPHICS Plus ...
Microsoft Excel - Libro1
06:57 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [Hypothesis Tests]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Hypothesis Tests
Sample mean = 3.0
Sample standard deviation = 0.71
Sample size = 38
95.0% confidence interval for mean: 3.
Null Hypothesis: mean = 2.0
Alternative: not equal
Computed t statistic = 8.68227
P-Value = 1.86277E-10
Reject the null hypothesis for alpha =
The StatAdvisor
This analysis shows the results of
concerning the mean (mu) of a normal
to be tested are:
Null hypothesis: mu = 2.0
Alternative hypothesis: mu <> 2.0
Given a sample of 38 observations with a mean of 3.0 and a standard
deviation of 0.71, the computed t statistic equals 8.68227. Since the
P-value for the test is less than 0.05, the null hypothesis is
rejected at the 95.0% confidence level. The confidence interval shows
that the values of mu supported by the data fall between 2.76663 and
3.23337.
Hypothesis Tests Options
Alternative Hypothesis
Not Equal
Less Than
Greater Than
OK
Cancel
Help
Alpha:
1
%
Ready
NUM
Inicio
Documento1 - Micros...
STATGRAPHICS Plus ...
Microsoft Excel - Libro1
06:58 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [Hypothesis Tests]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Hypothesis Tests
Sample mean = 3.0
Sample standard deviation = 0.71
Sample size = 38
99.0% lower confidence bound for mean: 3.0 - 0.280048 [2.71995]
Null Hypothesis: mean = 2.0
Alternative: greater than
Computed t statistic = 8.68227
P-Value = 9.31384E-11
Reject the null hypothesis for alpha = 0.01.
The StatAdvisor
This analysis shows the results of performing a hypothesis test
concerning the mean (mu) of a normal distribution. The two hypotheses
to be tested are:
Null hypothesis: mu = 2.0
Alternative hypothesis: mu > 2.0
Given a sample of 38 observations with a mean of 3.0 and a standard
deviation of 0.71, the computed t statistic equals 8.68227. Since the
P-value for the test is less than 0.01, the null hypothesis is
rejected at the 99.0% confidence level. The confidence bound shows
that the values of mu supported by the data are greater than or equal
to 2.71995.
Use the right mouse button to select options
NUM
Inicio
Documento1 - Micros...
STATGRAPHICS Plus ...
Microsoft Excel - Libro1
07:04 p.m.

Según se observa, el valor de probabilidad de la dócima es igual a 9.31^{-11}, o sea, un valor extremadamente pequeño, muy cercano a 0. Como ese valor de probabilidad es menor que 0.01 (nivel de significación de la dócima), entonces se cumple la región crítica y se rechaza la hipótesis nula. Finalmente se puede afirmar, que la sospecha de los estudiantes es cierta, pues la media de la duración de los sub-procesos de Alimentos y Bebidas en el hotel, está superando los 2 minutos con un nivel de significación del 1%.

4.3. Dócima de hipótesis de la desviación típica.

Véase un ejemplo.

Ejemplo 2:

Se conoce que la desviación típica de los ingresos de un restaurante de la red extrahotelera en Cienfuegos, es de 250 pesos. Hace unas semanas, abrieron un nuevo punto de venta cerca del restaurante, lo cual constituye una amenaza para los directivos de este último, respecto a los niveles de ingresos. Para conocer si la variabilidad de los ingresos ha continuado como antes, la dirección del restaurante tomó una muestra de 18 días y detectó que la desviación estándar de los ingresos, fue de 186 pesos con un nivel de confiabilidad del 95%.

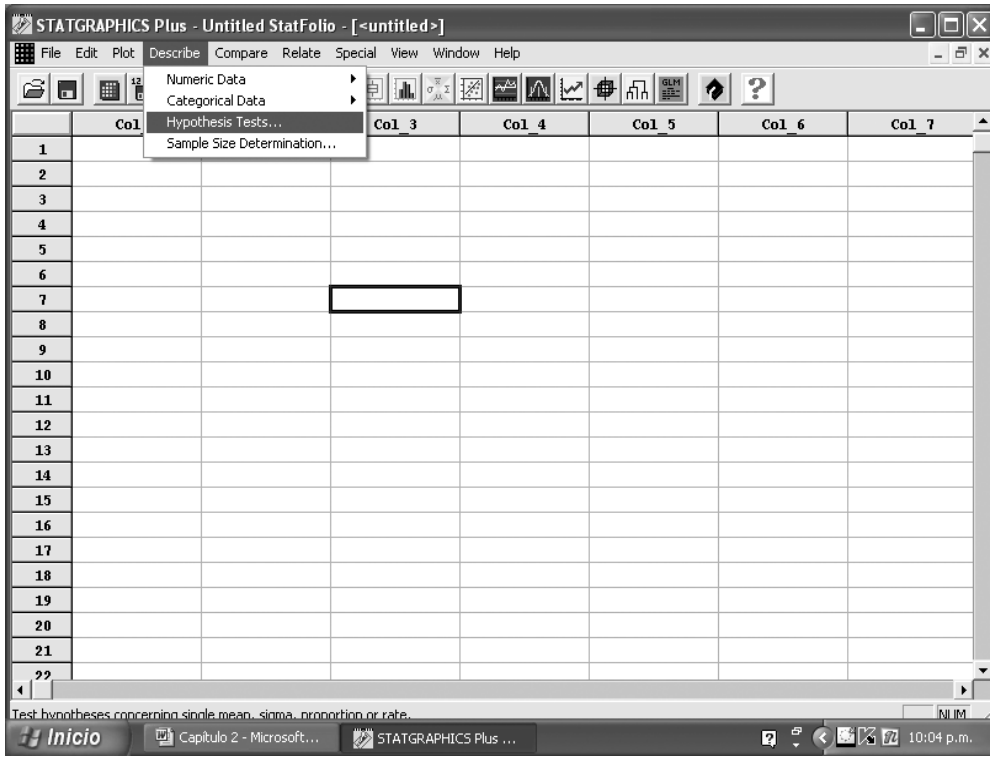
Solución:

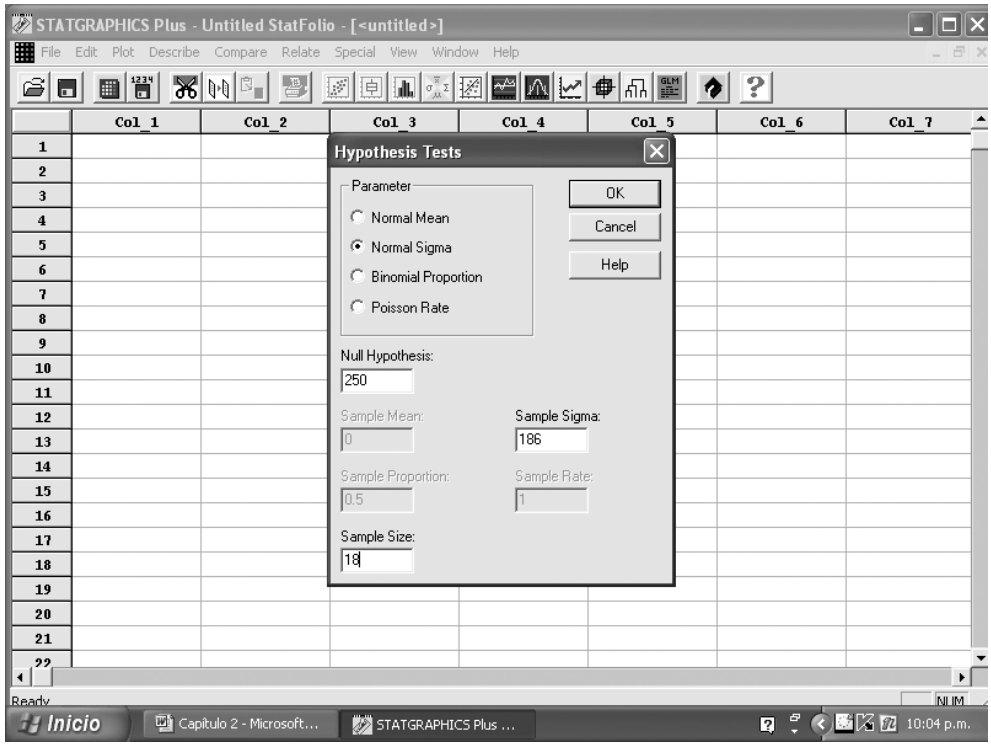
Véase que este es un ejemplo donde se realizará una dócima de hipótesis acerca de la **desviación típica** de la variable "ingresos de un restaurante de la red extrahotelera".

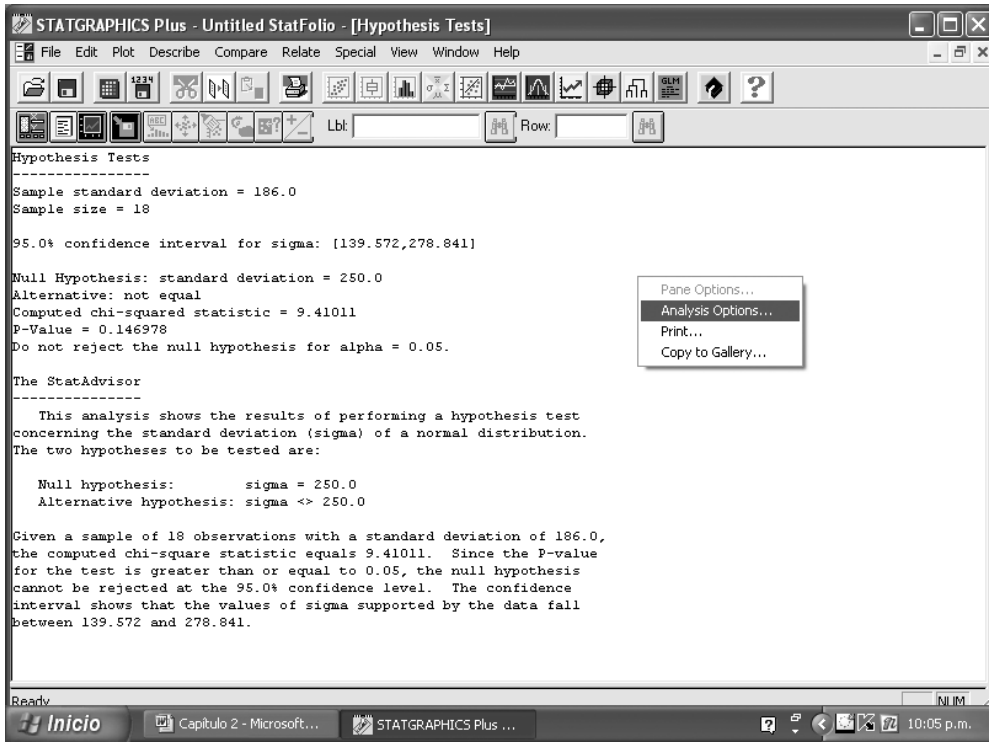
H_0: $\sigma = 250$ (desviación estándar igual a \$250)

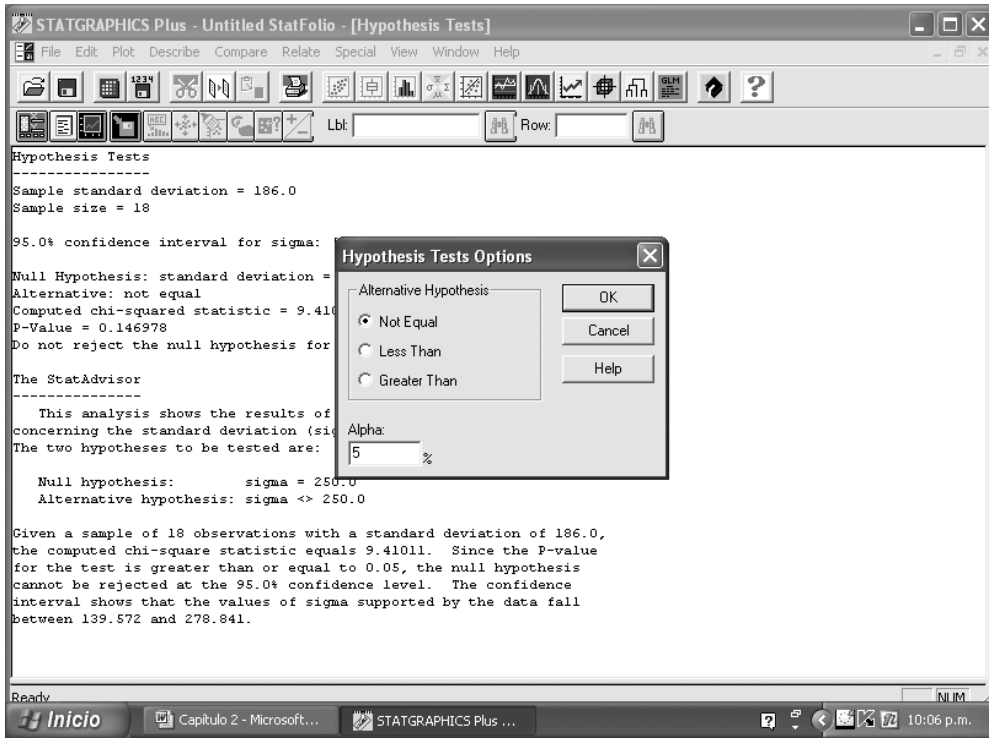
H_1: $\sigma \neq 250$ (desviación estándar diferente a \$250)

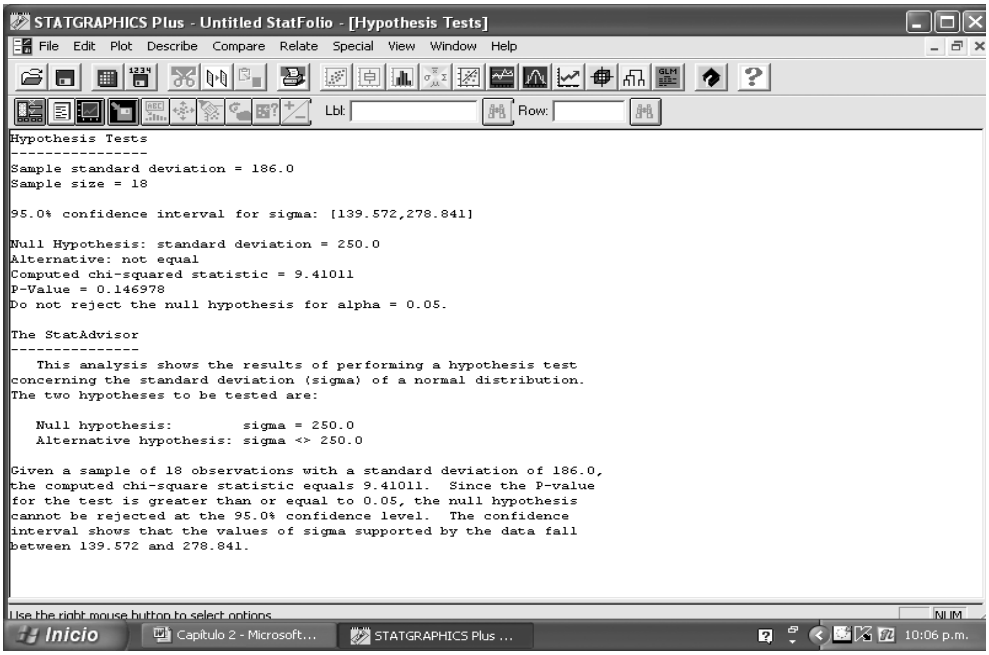
Empleando el STATGRAPHICS, sería:


STATGRAPHICS Plus - Untitled StatFolio - [<untitled>]
File Edit Plot Describe Compare Relate Special View Window Help
Numeric Data
Categorical Data
Hypothesis Tests...
Sample Size Determination...
Col_1 Col_2 Col_3 Col_4 Col_5 Col_6 Col_7
Test hypotheses concerning single mean, sigma, proportion or rate.
Inicio
Capítulo 2 - Microsoft...
STATGRAPHICS Plus ...
10:04 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [<untitled>]
Hypothesis Tests
Parameter
Normal Mean
Normal Sigma
Binomial Proportion
Poisson Rate
OK
Cancel
Help
Null Hypothesis:
250
Sample Mean:
0
Sample Sigma:
186
Sample Proportion:
0.5
Sample Rate:
1
Sample Size:
18
Ready
Inicio
Capítulo 2 - Microsoft...
STATGRAPHICS Plus ...
10:04 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [Hypothesis Tests]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Hypothesis Tests
Sample standard deviation = 186.0
Sample size = 18
95.0% confidence interval for sigma: [139.572,278.841]
Null Hypothesis: standard deviation = 250.0
Alternative: not equal
Computed chi-squared statistic = 9.41011
P-Value = 0.146978
Do not reject the null hypothesis for alpha = 0.05.
The StatAdvisor
This analysis shows the results of performing a hypothesis test
concerning the standard deviation (sigma) of a normal distribution.
The two hypotheses to be tested are:
Null hypothesis: sigma = 250.0
Alternative hypothesis: sigma <> 250.0
Given a sample of 18 observations with a standard deviation of 186.0,
the computed chi-square statistic equals 9.41011. Since the P-value
for the test is greater than or equal to 0.05, the null hypothesis
cannot be rejected at the 95.0% confidence level. The confidence
interval shows that the values of sigma supported by the data fall
between 139.572 and 278.841.
Pane Options...
Analysis Options...
Print...
Copy to Gallery...
Ready
NUM
Inicio
Capítulo 2 - Microsoft...
STATGRAPHICS Plus ...
10:05 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [Hypothesis Tests]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Hypothesis Tests
Sample standard deviation = 186.0
Sample size = 18
95.0% confidence interval for sigma:
Null Hypothesis: standard deviation =
Alternative: not equal
Computed chi-squared statistic = 9.41
P-Value = 0.146978
Do not reject the null hypothesis for
The StatAdvisor
This analysis shows the results of
concerning the standard deviation (si
The two hypotheses to be tested are:
Null hypothesis: sigma = 250.0
Alternative hypothesis: sigma <> 250.0
Given a sample of 18 observations with a standard deviation of 186.0,
the computed chi-square statistic equals 9.41011. Since the P-value
for the test is greater than or equal to 0.05, the null hypothesis
cannot be rejected at the 95.0% confidence level. The confidence
interval shows that the values of sigma supported by the data fall
between 139.572 and 278.841.
Hypothesis Tests Options
Alternative Hypothesis
Not Equal
Less Than
Greater Than
OK
Cancel
Help
Alpha:
5
%
Ready
NUM
Inicio
Capítulo 2 - Microsoft...
STATGRAPHICS Plus ...
10:06 p.m.



Según se observa, el valor de probabilidad de la dócima es igual a 0.15. Como este valor no es menor que 0.05, entonces se acepta la hipótesis nula, lo cual significa que la dirección del restaurante ha podido comprobar que, a pesar de haber un nuevo punto de venta cerca, esto no ha provocado cambios en la desviación típica habitual de los ingresos con un nivel de significación del 5%.

4.4. Dócima de hipótesis de la proporción.

Véase un ejemplo.

Ejemplo 3:

Según datos históricos de una agencia de viajes, como mínimo el 58% de los clientes canadienses que compran excursiones en un año, se inclinan por la oferta "Habana Colonial". En el actual año, el Departamento de Comercial de la agencia, sospecha que los canadienses se están inclinando preferentemente hacia la compra de la excursión "Guamá", y para corroborar la hipótesis, el equipo de comerciales determinó que durante 43 días, el 62% de los clientes se ha inclinado por la compra de la excursión "Habana Colonial" con un nivel del confiabilidad del 90%.

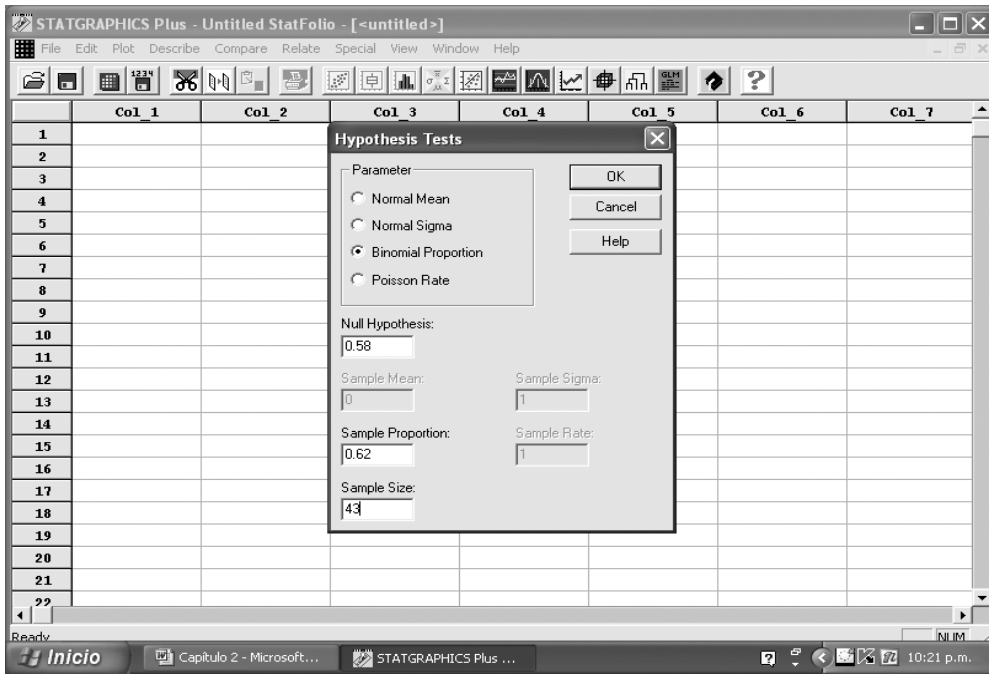
Solución:

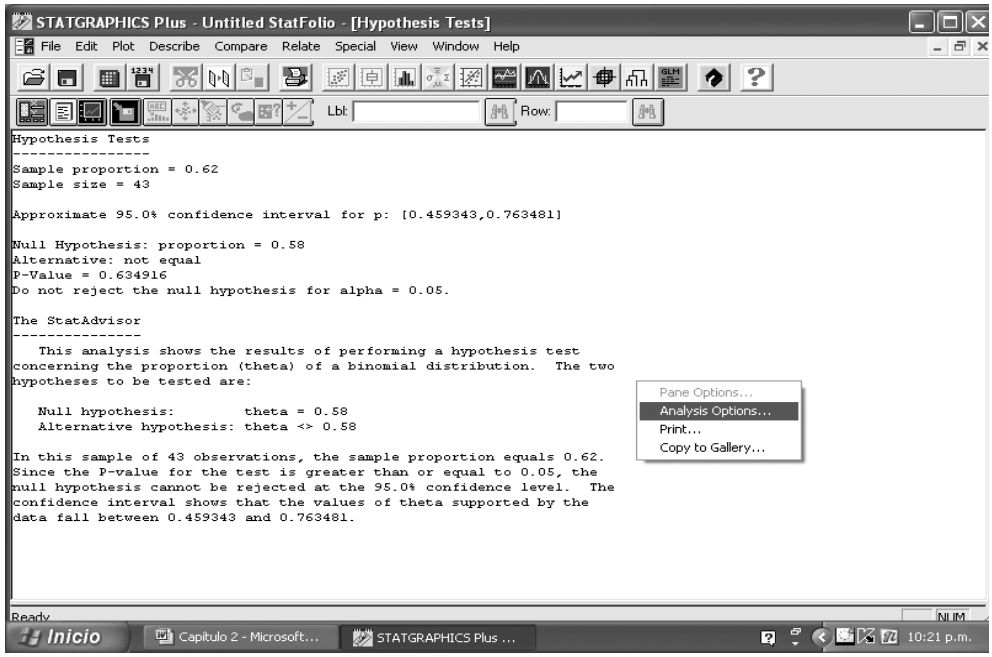
Véase que este es un ejemplo donde se realizará una dócima de hipótesis acerca de la **proporción** de la variable "cantidad de turistas que prefieren comprar la

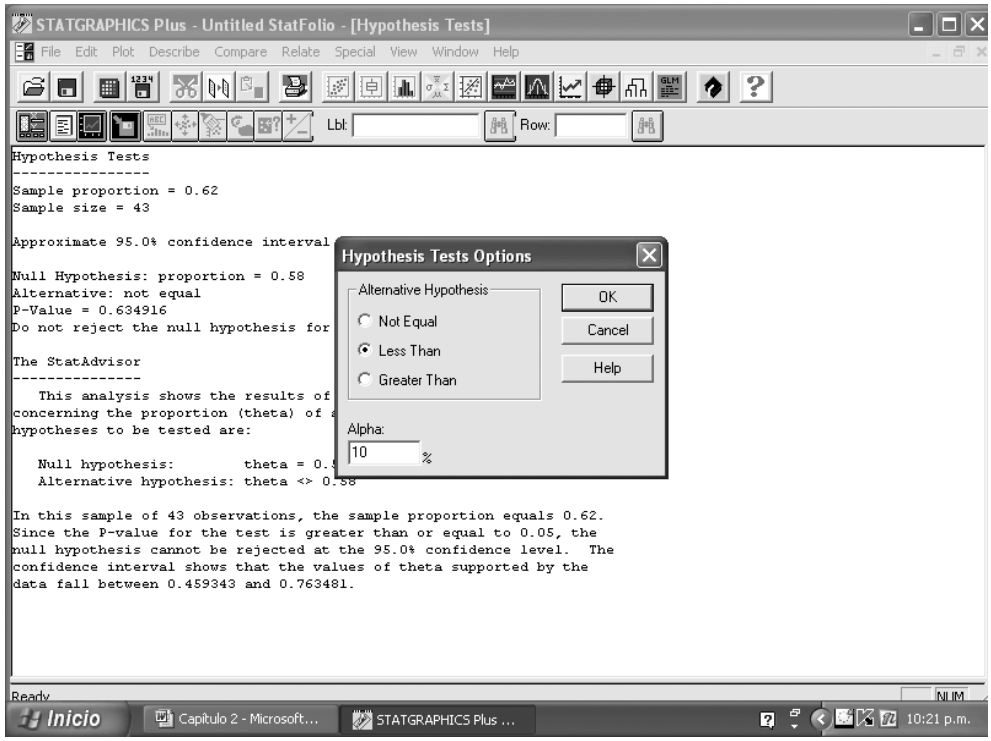
excursión Habana Colonial".

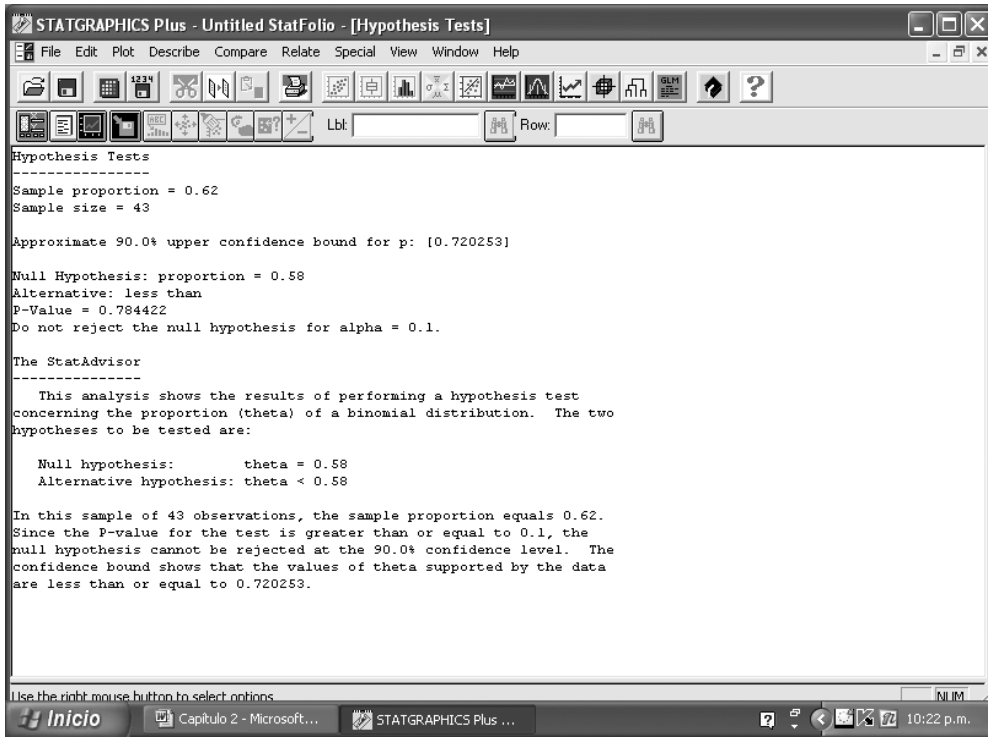
H_0: $p \geq 0.58$ (proporción igual o mayor que 58%)

H_1: $p < 0.58$ (proporción inferior a 58%)






STATGRAPHICS Plus - Untitled StatFolio - [Hypothesis Tests]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Hypothesis Tests
Sample proportion = 0.62
Sample size = 43
Approximate 95.0% confidence interval
Null Hypothesis: proportion = 0.58
Alternative: not equal
P-Value = 0.634916
Do not reject the null hypothesis for
The StatAdvisor
This analysis shows the results of
concerning the proportion (theta) of
hypotheses to be tested are:
Null hypothesis: theta = 0.
Alternative hypothesis: theta <> 0.58
In this sample of 43 observations, the sample proportion equals 0.62.
Since the P-value for the test is greater than or equal to 0.05, the
null hypothesis cannot be rejected at the 95.0% confidence level. The
confidence interval shows that the values of theta supported by the
data fall between 0.459343 and 0.763481.
Hypothesis Tests Options
Alternative Hypothesis
Not Equal
Less Than
Greater Than
OK
Cancel
Help
Alpha:
10
%
Ready
NUM
Inicio
Capítulo 2 - Microsoft...
STATGRAPHICS Plus ...
10:21 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [Hypothesis Tests]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Hypothesis Tests
Sample proportion = 0.62
Sample size = 43
Approximate 90.0% upper confidence bound for p: [0.720253]
Null Hypothesis: proportion = 0.58
Alternative: less than
P-Value = 0.784422
Do not reject the null hypothesis for alpha = 0.1.
The StatAdvisor
This analysis shows the results of performing a hypothesis test
concerning the proportion (theta) of a binomial distribution. The two
hypotheses to be tested are:
Null hypothesis: theta = 0.58
Alternative hypothesis: theta < 0.58
In this sample of 43 observations, the sample proportion equals 0.62.
Since the P-value for the test is greater than or equal to 0.1, the
null hypothesis cannot be rejected at the 90.0% confidence level. The
confidence bound shows that the values of theta supported by the data
are less than or equal to 0.720253.
Use the right mouse button to select options
NUM
Inicio
Capítulo 2 - Microsoft...
STATGRAPHICS Plus ...
10:22 p.m.

Según la información que ofrece la imagen anterior, el valor de probabilidad de la dócima es igual a 0.78 el cual no es menor que 0.10, por tanto, no se cumple la región crítica y se acepta la hipótesis nula. Se puede decir entonces que, la sospecha del Departamento de Comercial puede no ser cierta con un nivel de seguridad del 90%, pues como mínimo, el 58% de los turistas ha continuado prefiriendo la excursión "Habana Colonial", tal y como muestran los hatos históricos.

4.5. Dócima de hipótesis de la diferencia de medias.

Véase un ejemplo.

Ejemplo 4:

El jefe general de Alimentos y Bebidas de dos hoteles que pertenecen a una misma cadena hotelera, conoce que la rotura y pérdida de artículos tales como vajilla (platos, tasas, etc.) y cristalería (copas, vasos, etc.) en los restaurantes y bares, provoca un gasto mayor en el primer hotel que en el segundo, pues el primer hotel es de categoría 5 estrellas y el segundo de 4 estrellas, contando con mayor volumen de operaciones, el primer hotel.

Últimamente, los gastos de dichos artículos de reposición, según reflejan los balances de explotación de ambos hoteles, han sido mayores en el segundo, lo cual contradice el comportamiento comparativo habitual de esta cuenta o partida de gastos. Para verificar si la tendencia en los gastos comparados entre ambos hoteles ha variado, el jefe tomó una muestra de 26 días en el primer hotel, y verificó que el gasto promedio diario fue de 610 pesos con una varianza de 16 pesos^2. Por otro lado, estudió durante 24 días el comportamiento diario de los gastos en el segundo hotel, y comprobó que la media fue de 598 pesos con una desviación estándar de 5 pesos. Todo ello fue realizado con un nivel de confiabilidad del 95%.

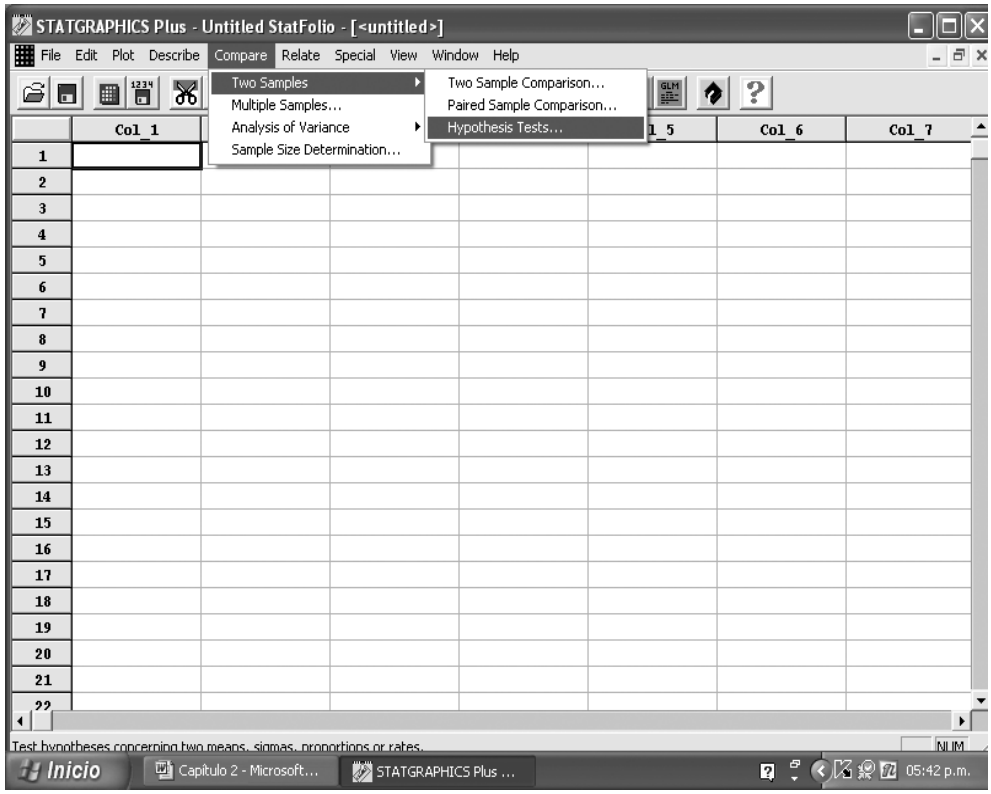
Solución:

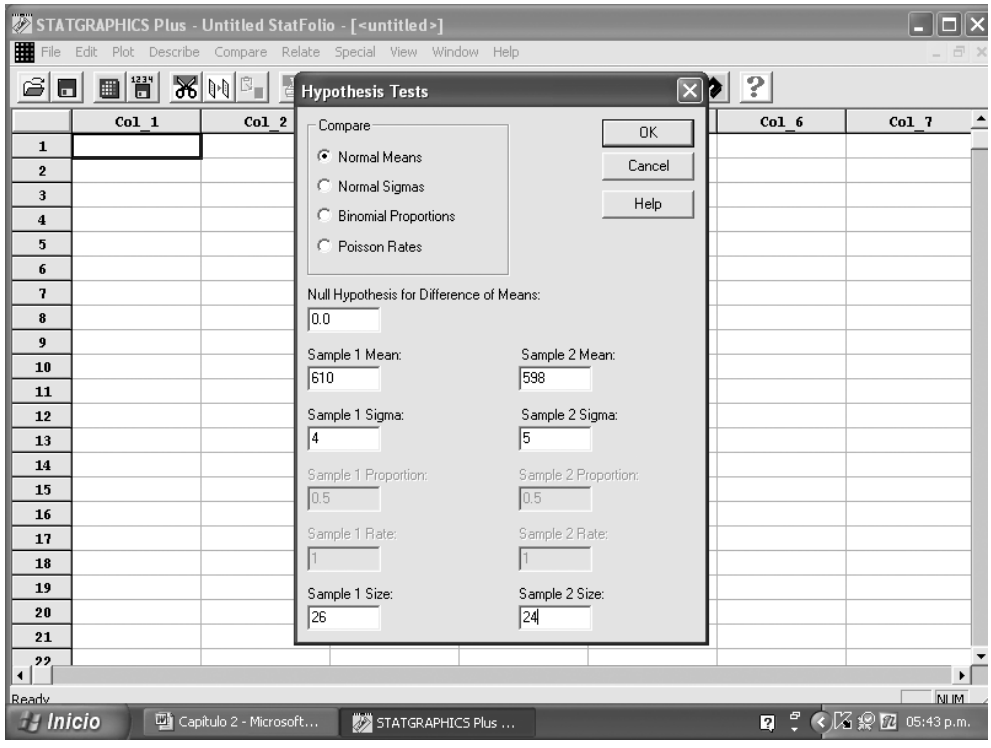
Véase que este es un ejemplo donde se realizará una dócima de hipótesis acerca de la **diferencia de medias** de la variable "gastos diarios de artículos de reposición".

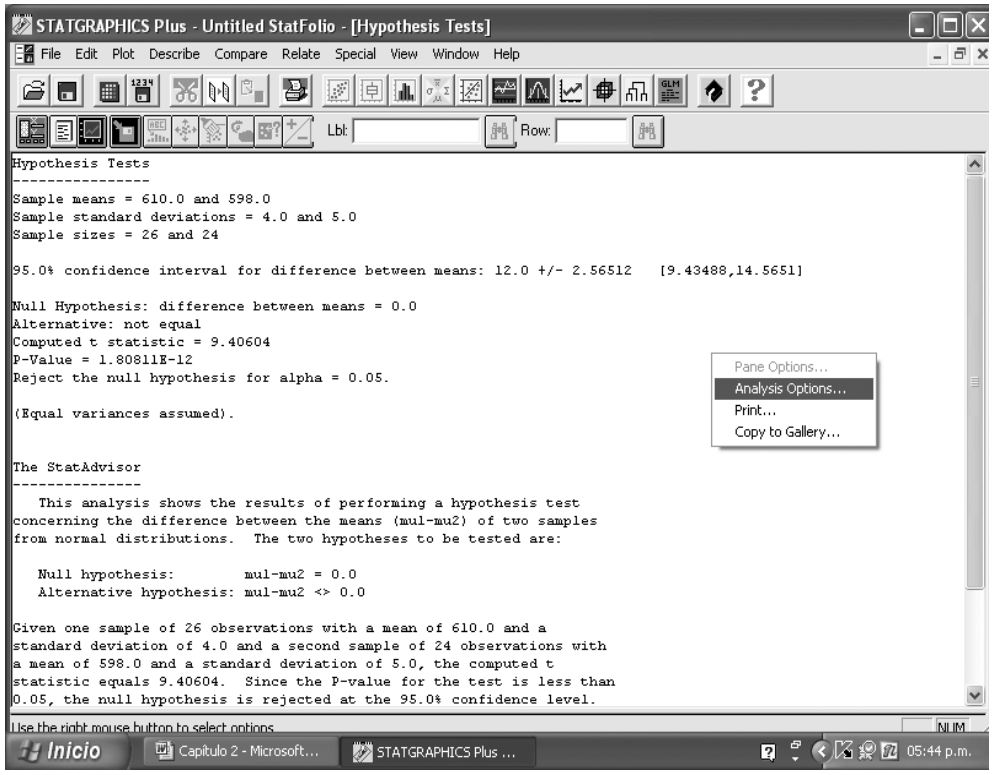
H_0: $\mu_1 \leq \mu_2$ (el promedio de gastos del primer hotel es menor o igual que el promedio de gastos del segundo hotel)

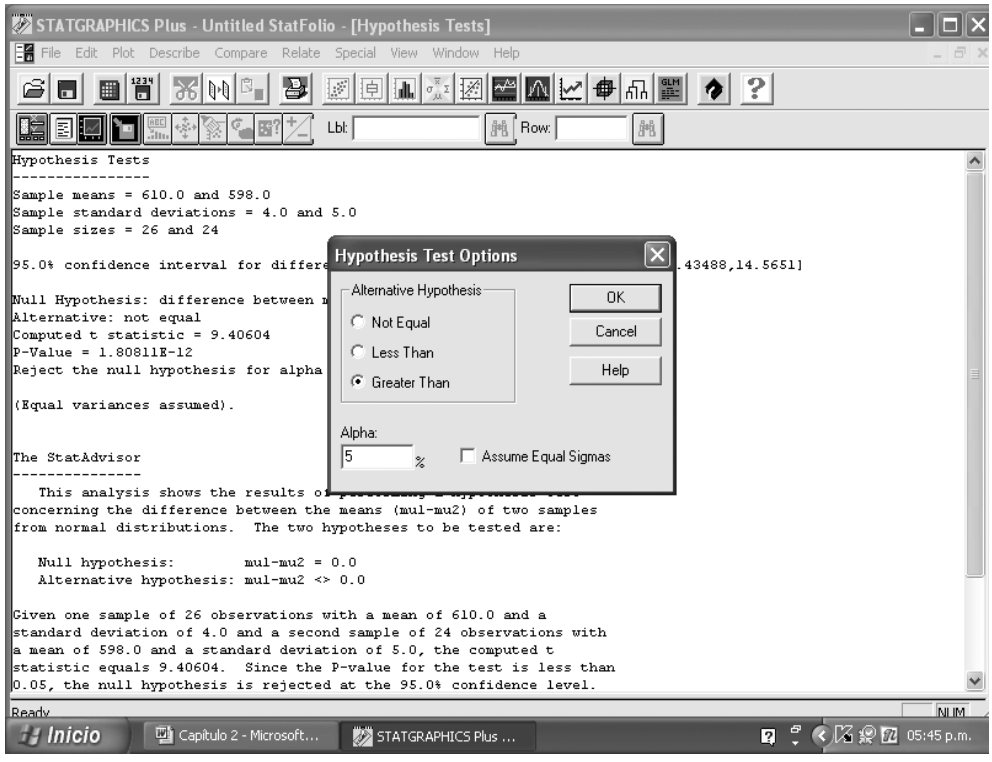
H_1: $\mu_1 > \mu_2$ (el promedio de gastos del primer hotel es superior al promedio de gastos del segundo hotel)

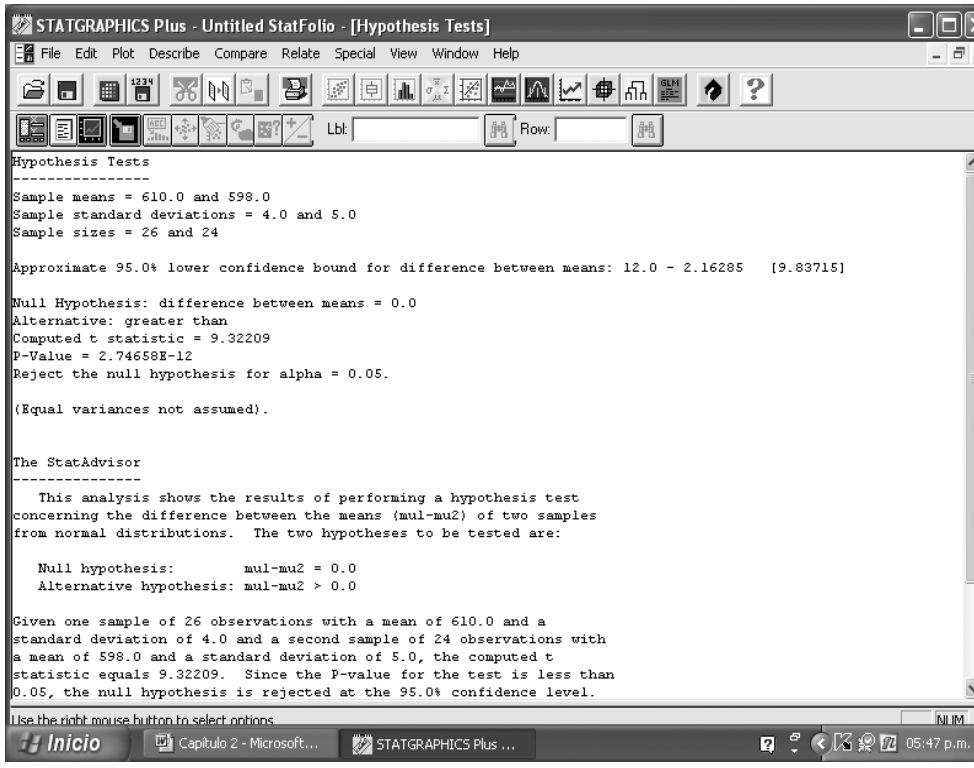
En el STATGRAPHICS se opera de la siguiente forma:


STATGRAPHICS Plus - Untitled StatFolio - [<untitled>]
File Edit Plot Describe Compare Relate Special View Window Help
Two Samples
Multiple Samples...
Analysis of Variance
Sample Size Determination...
Two Sample Comparison...
Paired Sample Comparison...
Hypothesis Tests...
Col_1
Col_6
Col_7
Test hypotheses concerning two means, sigmas, proportions or rates.
Inicio
Capitulo 2 - Microsoft...
STATGRAPHICS Plus ...
05:42 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [<untitled>]
Hypothesis Tests
Compare
Normal Means
Normal Sigmas
Binomial Proportions
Poisson Rates
OK
Cancel
Help
Null Hypothesis for Difference of Means:
0.0
Sample 1 Mean:
610
Sample 2 Mean:
598
Sample 1 Sigma:
4
Sample 2 Sigma:
5
Sample 1 Proportion:
0.5
Sample 2 Proportion:
0.5
Sample 1 Rate:
1
Sample 2 Rate:
1
Sample 1 Size:
26
Sample 2 Size:
24
Ready
Inicio
Capitulo 2 - Microsoft...
STATGRAPHICS Plus ...
05:43 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [Hypothesis Tests]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Hypothesis Tests
Sample means = 610.0 and 598.0
Sample standard deviations = 4.0 and 5.0
Sample sizes = 26 and 24
95.0% confidence interval for difference between means: 12.0 +/- 2.56512 [9.43488,14.5651]
Null Hypothesis: difference between means = 0.0
Alternative: not equal
Computed t statistic = 9.40604
P-Value = 1.80811E-12
Reject the null hypothesis for alpha = 0.05.
(Equal variances assumed).
The StatAdvisor
This analysis shows the results of performing a hypothesis test concerning the difference between the means (mul-mu2) of two samples from normal distributions. The two hypotheses to be tested are:
Null hypothesis: mul-mu2 = 0.0
Alternative hypothesis: mul-mu2 <> 0.0
Given one sample of 26 observations with a mean of 610.0 and a standard deviation of 4.0 and a second sample of 24 observations with a mean of 598.0 and a standard deviation of 5.0, the computed t statistic equals 9.40604. Since the P-value for the test is less than 0.05, the null hypothesis is rejected at the 95.0% confidence level.
Pane Options...
Analysis Options...
Print...
Copy to Gallery...
Use the right mouse button to select options
NUM
Inicio
Capítulo 2 - Microsoft...
STATGRAPHICS Plus ...
05:44 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [Hypothesis Tests]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Hypothesis Tests
Sample means = 610.0 and 598.0
Sample standard deviations = 4.0 and 5.0
Sample sizes = 26 and 24
Hypothesis Test Options
Alternative Hypothesis
Not Equal
Less Than
Greater Than
OK
Cancel
Help
Alpha:
5
%
Assume Equal Sigmas
Null Hypothesis: difference between
Alternative: not equal
Computed t statistic = 9.40604
P-Value = 1.80811E-12
Reject the null hypothesis for alpha
(Equal variances assumed).
The StatAdvisor
concerning the difference between the means (mul-mu2) of two samples from normal distributions. The two hypotheses to be tested are:
Null hypothesis: mul-mu2 = 0.0
Alternative hypothesis: mul-mu2 <> 0.0
Given one sample of 26 observations with a mean of 610.0 and a standard deviation of 4.0 and a second sample of 24 observations with a mean of 598.0 and a standard deviation of 5.0, the computed t statistic equals 9.40604. Since the P-value for the test is less than 0.05, the null hypothesis is rejected at the 95.0% confidence level.
Ready
NUM
Inicio
Capítulo 2 - Microsoft...
STATGRAPHICS Plus ...
05:45 p.m.



Puede observarse que el valor de probabilidad de la dócima es muy pequeño, igual a 2.75^{-12}. Como este valor es menor que 0.05, entonces se cumple la región crítica y se rechaza la hipótesis nula, por lo cual puede afirmarse que los gastos del primer hotel, continúan siendo mayores que los del segundo, tal y como se ha comportado habitualmente el nivel de operaciones, con un nivel del significación del 5%.

4.6. Dócima de hipótesis de la diferencia de proporciones.

Véase un ejemplo.

Ejemplo 5:

La Delegación del Turismo de un territorio, opina que el porcentaje de trabajadores de un Hotel X que dice conocer el concepto de calidad en los servicios con que trabaja el mismo, es similar al porcentaje en el Hotel V que dice conocer el concepto con que labora este otro. Para deshacer las dudas, la Delegación aplicó un pequeño cuestionario a los empleados de ambos hoteles, el cual se muestra a continuación:

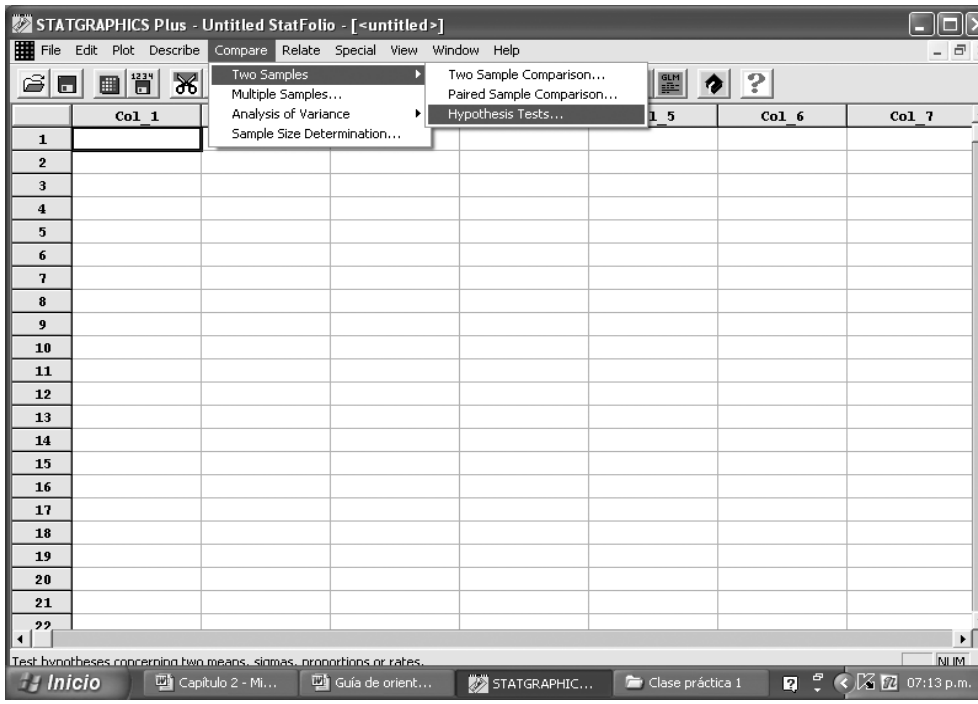
¿Conoce usted el concepto de calidad con que trabaja su entidad?
Sí: __ No: __
Si marcó la opción "sí", por favor, enuncie dicho concepto.

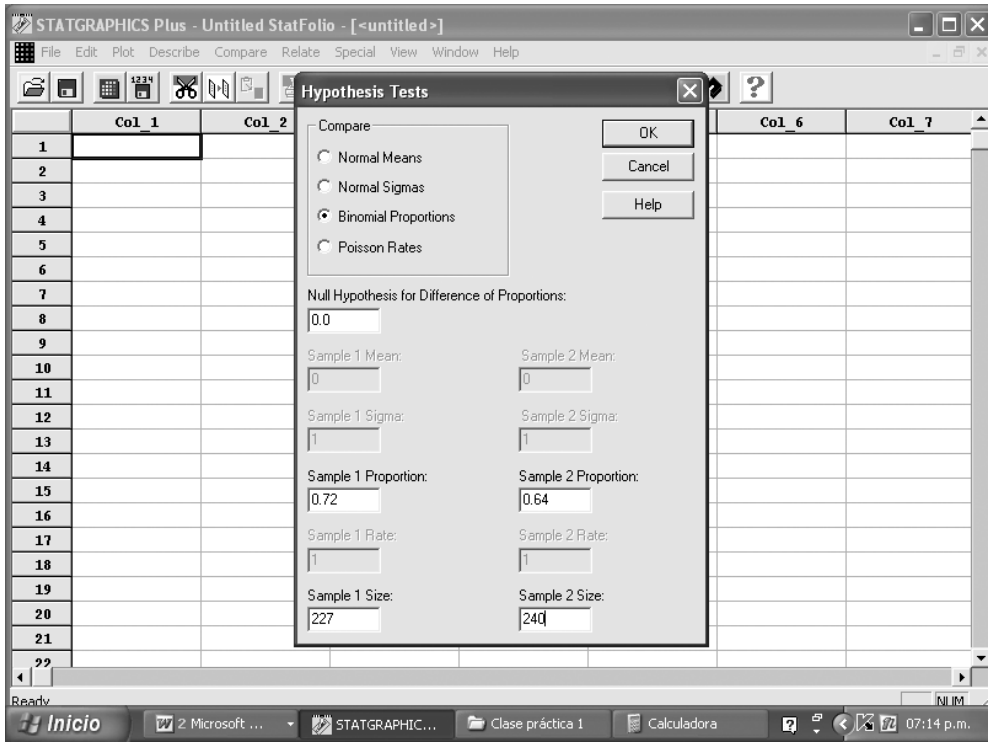
Después de realizada la encuesta, se comprobó que 163 trabajadores de un total de 227 en el Hotel X, dijeron que sí conocían el concepto de calidad en los servicios con que laboraba la entidad. Por su parte, 154 empleados de 240 en el Hotel V, plantearon que sí conocían igualmente el concepto con que trabajaba la instalación. La Delegación desea entonces demostrar, si existen o no diferencias significativas entre las proporciones de trabajadores en cada hotel, que conocen el concepto de calidad con que labora cada uno, con un nivel de seguridad del 99%.
Solución:

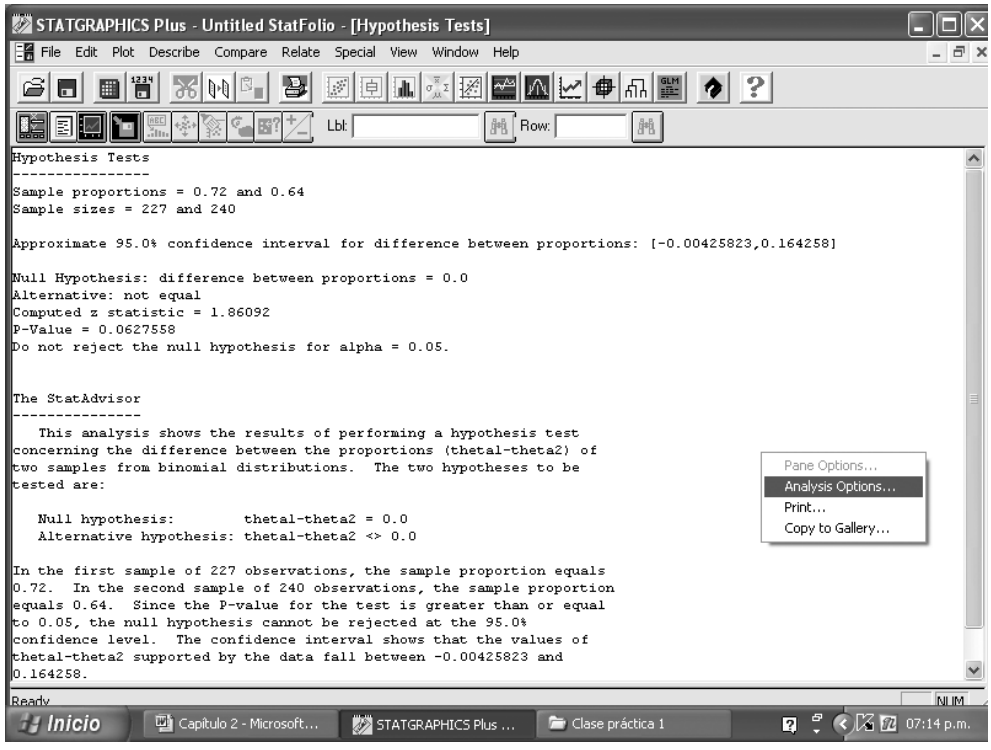
Véase que este es un ejemplo donde se realizará una dócima de hipótesis acerca de la **diferencia de proporciones** de la variable "conocer el concepto de calidad con que laboran dos hoteles".

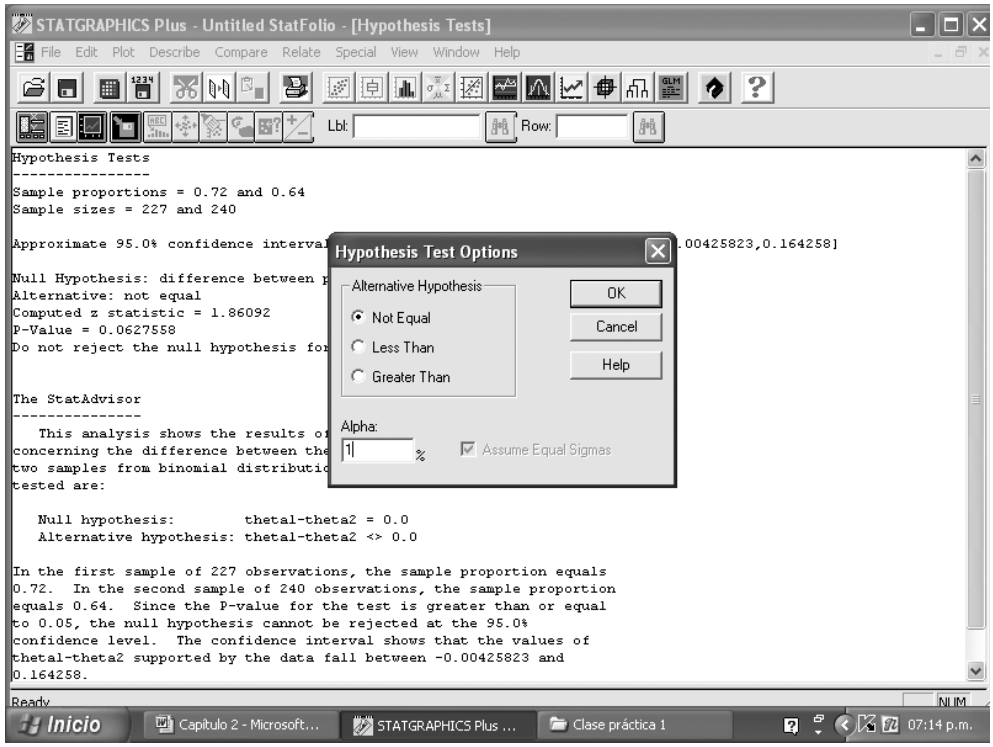
H_0: $p_1 = p_2$ (la proporción de trabajadores que sí conoce el concepto de calidad, es igual en ambos hoteles)

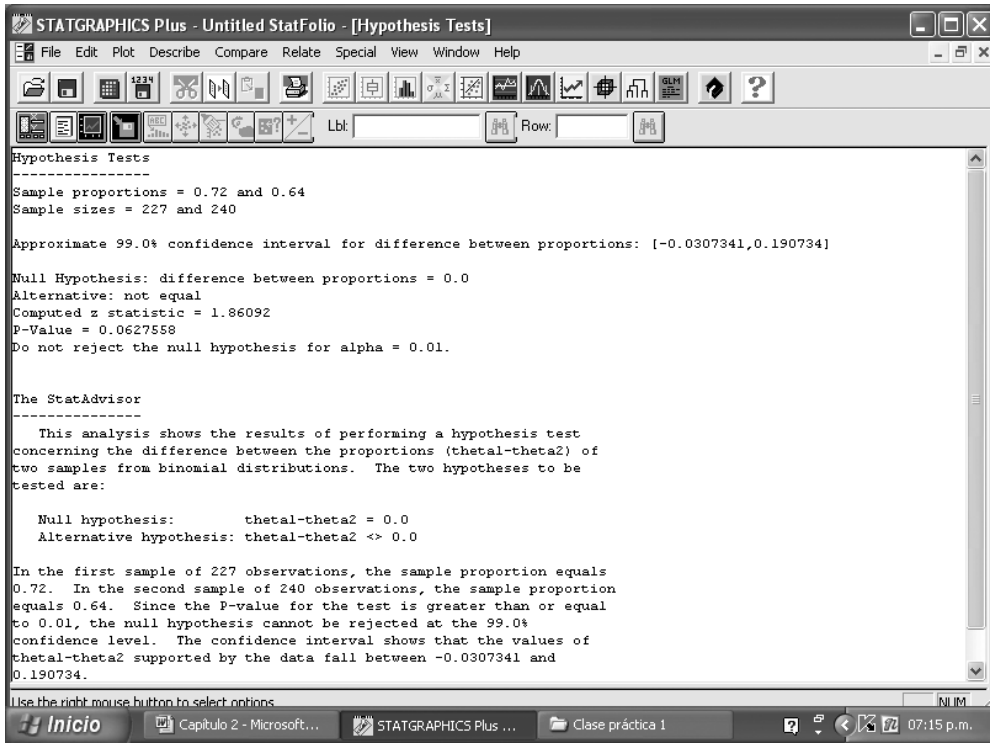
H_1: $p_1 \neq p_2$ (la proporción de trabajadores que sí conoce el concepto de calidad, es diferente en ambos hoteles)




STATGRAPHICS Plus - Untitled StatFolio - [<untitled>]
File Edit Plot Describe Compare Relate Special View Window Help
Hypothesis Tests
Compare
Normal Means
Normal Sigmas
Binomial Proportions
Poisson Rates
OK
Cancel
Help
Null Hypothesis for Difference of Proportions:
0.0
Sample 1 Mean:
Sample 2 Mean:
Sample 1 Sigma:
Sample 2 Sigma:
Sample 1 Proportion:
0.72
Sample 2 Proportion:
0.64
Sample 1 Rate:
Sample 2 Rate:
Sample 1 Size:
227
Sample 2 Size:
240
Col_1
Col_2
Col_6
Col_7
Ready
NUM
Inicio
2 Microsoft ...
STATGRAPHIC...
Clase práctica 1
Calculadora
07:14 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [Hypothesis Tests]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl:
Row:
Hypothesis Tests
Sample proportions = 0.72 and 0.64
Sample sizes = 227 and 240
Approximate 95.0% confidence interval for difference between proportions: [-0.00425823,0.164258]
Null Hypothesis: difference between proportions = 0.0
Alternative: not equal
Computed z statistic = 1.86092
P-Value = 0.0627558
Do not reject the null hypothesis for alpha = 0.05.
The StatAdvisor
This analysis shows the results of performing a hypothesis test concerning the difference between the proportions (thetal-theta2) of two samples from binomial distributions. The two hypotheses to be tested are:
Null hypothesis: thetal-theta2 = 0.0
Alternative hypothesis: thetal-theta2 <> 0.0
In the first sample of 227 observations, the sample proportion equals 0.72. In the second sample of 240 observations, the sample proportion equals 0.64. Since the P-value for the test is greater than or equal to 0.05, the null hypothesis cannot be rejected at the 95.0% confidence level. The confidence interval shows that the values of thetal-theta2 supported by the data fall between -0.00425823 and 0.164258.
Pane Options...
Analysis Options...
Print...
Copy to Gallery...
Ready
NUM
Inicio
Capítulo 2 - Microsoft...
STATGRAPHICS Plus ...
Clase práctica 1
07:14 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [Hypothesis Tests]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Hypothesis Tests
Sample proportions = 0.72 and 0.64
Sample sizes = 227 and 240
Approximate 95.0% confidence interval
.00425823,0.164258]
Null Hypothesis: difference between p
Alternative: not equal
Computed z statistic = 1.86092
P-Value = 0.0627558
Do not reject the null hypothesis for
The StatAdvisor
This analysis shows the results of
concerning the difference between the
two samples from binomial distributio
tested are:
Null hypothesis: thetal-theta2 = 0.0
Alternative hypothesis: thetal-theta2 <> 0.0
In the first sample of 227 observations, the sample proportion equals
0.72. In the second sample of 240 observations, the sample proportion
equals 0.64. Since the P-value for the test is greater than or equal
to 0.05, the null hypothesis cannot be rejected at the 95.0%
confidence level. The confidence interval shows that the values of
thetal-theta2 supported by the data fall between -0.00425823 and
0.164258.
Hypothesis Test Options
Alternative Hypothesis
Not Equal
Less Than
Greater Than
OK
Cancel
Help
Alpha:
1
%
Assume Equal Sigmas
Ready
NUM
Inicio
Capítulo 2 - Microsoft...
STATGRAPHICS Plus ...
Clase práctica 1
07:14 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [Hypothesis Tests]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Hypothesis Tests
Sample proportions = 0.72 and 0.64
Sample sizes = 227 and 240
Approximate 99.0% confidence interval for difference between proportions: [-0.0307341,0.190734]
Null Hypothesis: difference between proportions = 0.0
Alternative: not equal
Computed z statistic = 1.86092
P-Value = 0.0627558
Do not reject the null hypothesis for alpha = 0.01.
The StatAdvisor
This analysis shows the results of performing a hypothesis test
concerning the difference between the proportions (thetal-theta2) of
two samples from binomial distributions. The two hypotheses to be
tested are:
Null hypothesis: thetal-theta2 = 0.0
Alternative hypothesis: thetal-theta2 <> 0.0
In the first sample of 227 observations, the sample proportion equals
0.72. In the second sample of 240 observations, the sample proportion
equals 0.64. Since the P-value for the test is greater than or equal
to 0.01, the null hypothesis cannot be rejected at the 99.0%
confidence level. The confidence interval shows that the values of
thetal-theta2 supported by the data fall between -0.0307341 and
0.190734.
Use the right mouse button to select options
NUM
Inicio
Capítulo 2 - Microsoft...
STATGRAPHICS Plus ...
Clase práctica 1
07:15 p.m.

Como el valor de probabilidad de la dócima es igual a 0.06 y este no es menor que 0.01, entonces no se cumple la región crítica y se acepta la hipótesis nula, por lo cual la Delegación del Turismo en el territorio, pudiera valorar que no existan diferencias significativas entre las proporciones de trabajadores en ambos hoteles, que dicen conocer el concepto de calidad en los servicios con que laboran los mismos, todo ello bajo un nivel de significación del 1%.

EJERCITACIÓN

En la Transportista W, históricamente es asignado diariamente a las excursiones, circuitos y transfers de las agencias de viajes del polo turístico, un promedio de 65 ómnibus de 20, 30 y 40 plazas. En el mes de enero en curso, se ha elevado la cantidad de turistas hospedados en el polo, sin embargo, el director de la Transportista sospecha que no ha aumentado la venta de excursiones y opcionales por parte de las agencias de viajes, de modo que la cantidad de ómnibus empleados diariamente, no ha variado. Para deshacer las dudas, el director ha escogido una muestra de 18 días de este mes de enero, y ha comprobado que la cantidad promedio de ómnibus asignada diariamente a las agencias de viajes, ha sido de 68 buses con una varianza de 3 ómnibus^2. ¿Será cierta la sospecha del director de la Transportista con un nivel de significación del 5%?

SOLUCIÓN

H0: $\mu = 65$

H1: $\mu \neq 65$

Valor de probabilidad de la dócima: 0.00000111989

La sospecha del director de la Transportista W, pudiera no ser cierta con un nivel de confiabilidad del 95%, pues la cantidad promedio de ómnibus asignada diariamente en este mes de enero en curso a las agencias de viajes, no ha sido de 65 ómnibus, como venía ocurriendo de forma habitual, sino de más buses.

Capítulo
5

Pruebas de hipótesis no paramétricas.

5.1. Generalidades acerca de las dócimas de hipótesis no paramétricas.

Las dócimas de hipótesis no paramétricas se diferencian de las paramétricas en que estas pruebas no se basan en ninguna suposición en cuanto a la distribución de probabilidad a partir de la que fueron obtenidos los datos (distribution free).

5.2. Diversidad de pruebas de hipótesis no paramétricas.

Son muchas las pruebas de este tipo y que se emplean con diversos fines. A continuación, se comenta acerca de algunas más utilizadas:

- Prueba X^2 (Chi-Cuadrado) de Pearson: se emplea para verificar el ajuste de los datos a una distribución normal por lo cual también se le conoce como "prueba de bondad de ajuste"
- Prueba de Kolmogorov-Smirnov: se emplea para verificar el ajuste de los datos a una distribución normal. Es únicamente válida para variables continuas
- Prueba de Shapiro-Wilk: es la dócima que se recomienda para verificar el ajuste de los datos a una distribución normal sobre todo cuando la muestra es pequeña ($n < 30$). Brinda además una información gráfica que permite apreciar mejor el ajuste o desajuste

- Prueba de Wilcoxon: se emplea para verificar el ajuste de los datos a una distribución normal. Utiliza como parámetro de centralización, la mediana
- Prueba de Mann-Whitney: se emplea para comprobar la heterogeneidad de dos muestras ordinales
- Prueba de Kruskal-Wallis: constituye un análisis de varianza para escalas ordinales pero que no requiere condiciones de normalidad ni homocedasticidad de los datos. Sí exige aleatoriedad en la extracción de las muestras
- Prueba de Friedman: es una prueba equivalente a la ANOVA para dos factores, pero en su versión no paramétrica
- Prueba de Wald-Wolfowitz: también conocida como "prueba de rachas"
- Prueba de independencia X^2 empleando tablas de contingencia
- Prueba X^2 para determinar concordancia casual entre expertos

5.3. Prueba X^2 de Pearson (bondad de ajuste).

Véase un ejemplo.

Ejemplo 1:

El Departamento de Riesgo del Hotel S, está analizando de cada uno de sus principales turoperadores (TT.OO.), los días que demora en pagarle a la entidad. Para ello, el jefe del departamento ha recogido el nombre de dichos TT.OO. con el promedio de días en que cada uno demora en pagarle al hotel.

Con los datos que se muestran a continuación, el jefe desea saber si los promedios de días de pago, siguen o no una distribución normal con un nivel de confiabilidad del 99%.

TT.OO	Promedio de días de pago
Red Seal My Travel	35
Vacances Transat	32
Signature	28
Tour Mont Royal	36
Conquest Hola Sun Caribe Sol	34
FTI	23
All Tours	26
ITS	41
Thomas Cook UK	35
TUI Group	37
Special Traffic	42
Transnico International	24
Holiday Place	19
Kuoni Travel	34
Lotus Group	18
Travel COSAT	32
Travel Plan	27
Politour	26
Club Vacaciones	29
El Corte Inglés	31
Julia Tours	28
Free Way	18
Eves	24
Eurovip´s	31
Polimex	30
Taíno Tours	28
Tip´s	27
Sol y Son	26
Thomas Cook Alemania	32
First Choice	28
Sunwing	26
Friendship Tours	30

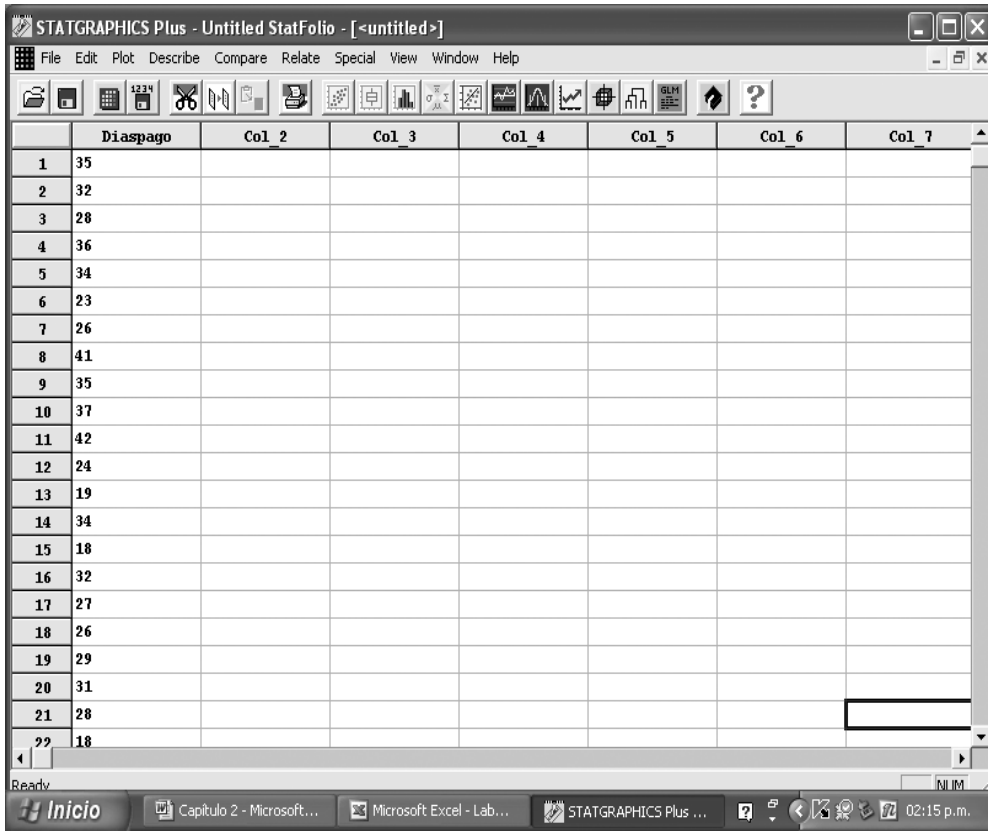
Solución:

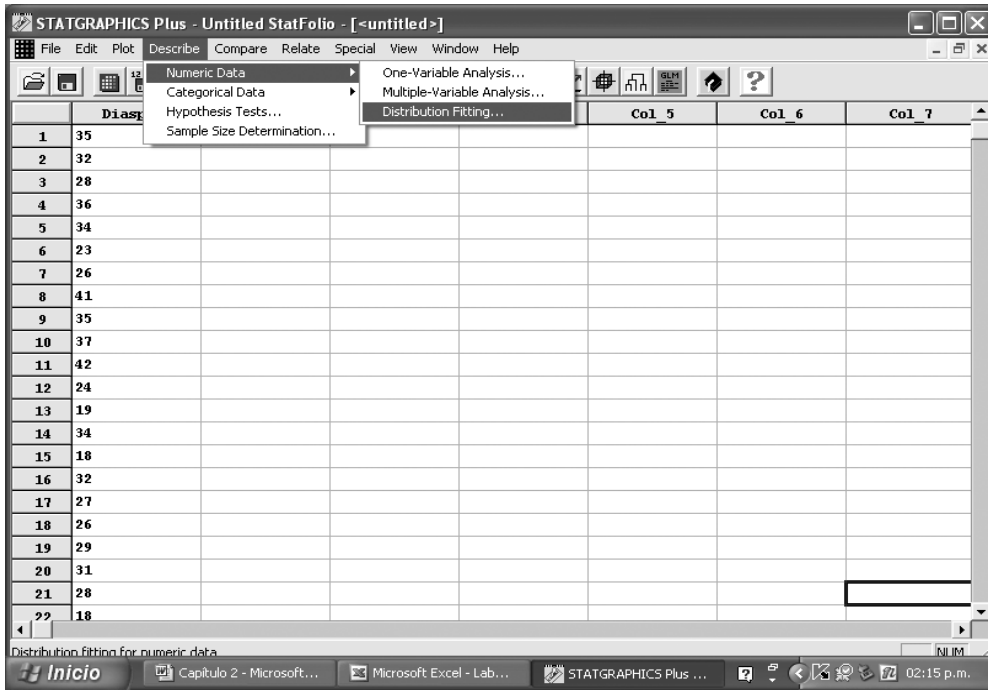
Variable de estudio X: días de pago.

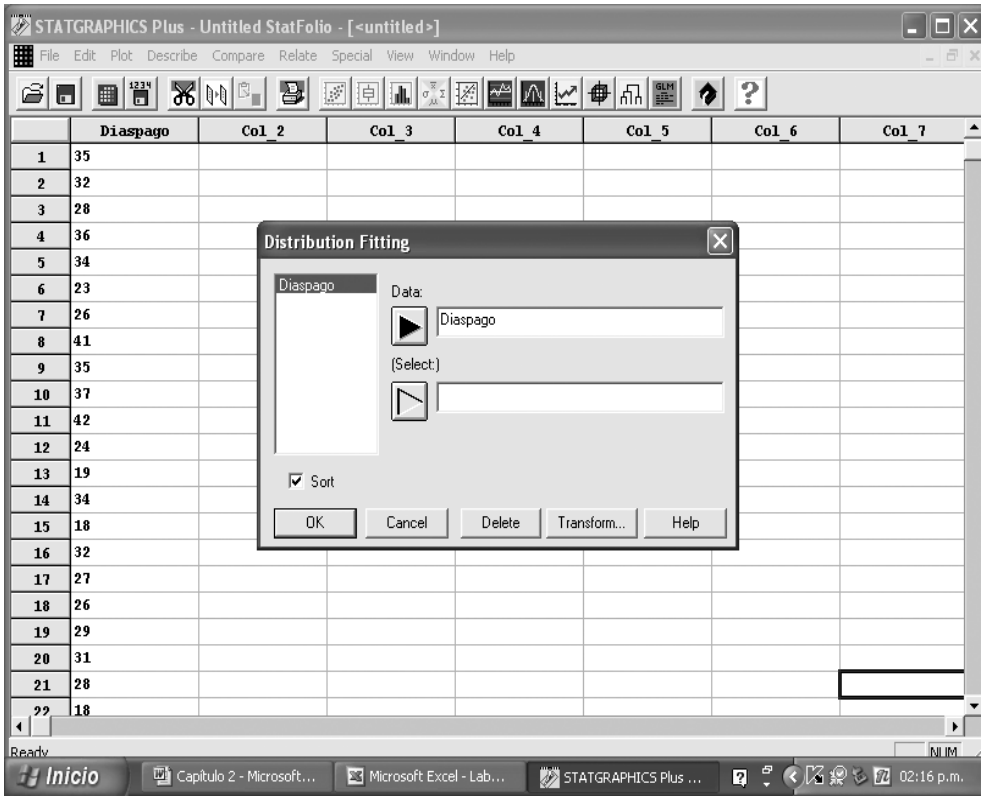
H_0: $X \sim N(\mu; \sigma^2)$ (la cantidad de días de pago sigue una distribución normal)

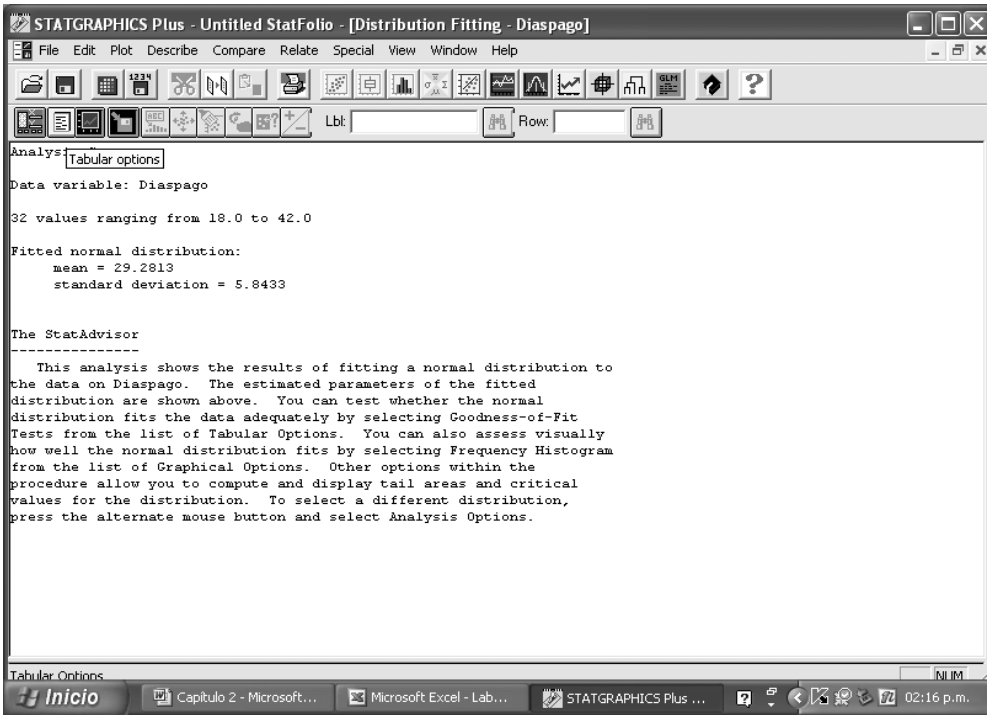
H_1: $X \overset{no}{\sim} N(\mu; \sigma^2)$ (la cantidad de días de pago, no sigue una distribución normal)

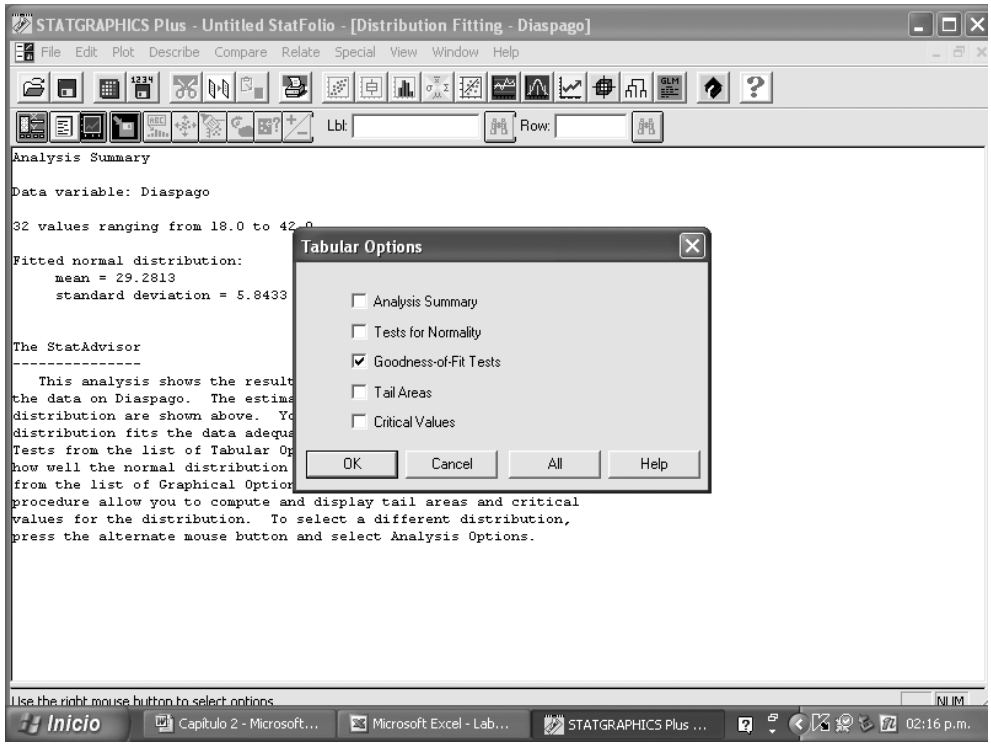
Utilizando el STATGRAPHICS Plus para realizar esta dócima de hipótesis no paramétrica, sería:

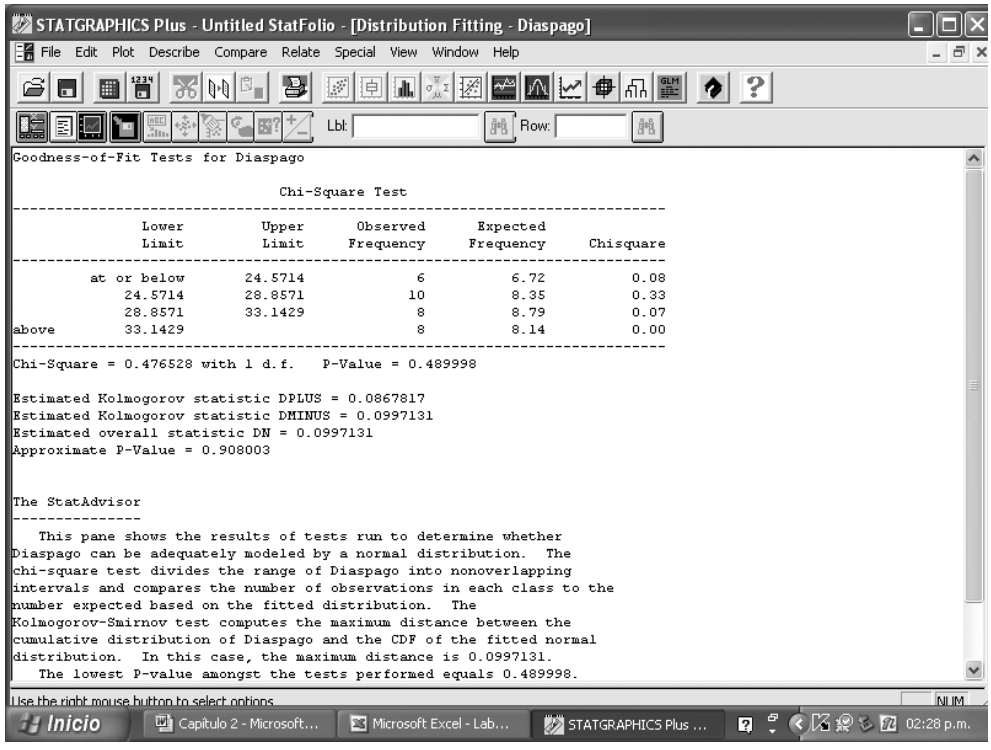

STATGRAPHICS Plus - Untitled StatFolio - [<untitled>]
File Edit Plot Describe Compare Relate Special View Window Help
Diaspago Col_2 Col_3 Col_4 Col_5 Col_6 Col_7
1 35
2 32
3 28
4 36
5 34
6 23
7 26
8 41
9 35
10 37
11 42
12 24
13 19
14 34
15 18
16 32
17 27
18 26
19 29
20 31
21 28
22 18
Ready
NUM
Inicio
Capítulo 2 - Microsoft...
Microsoft Excel - Lab...
STATGRAPHICS Plus ...
02:15 p.m.

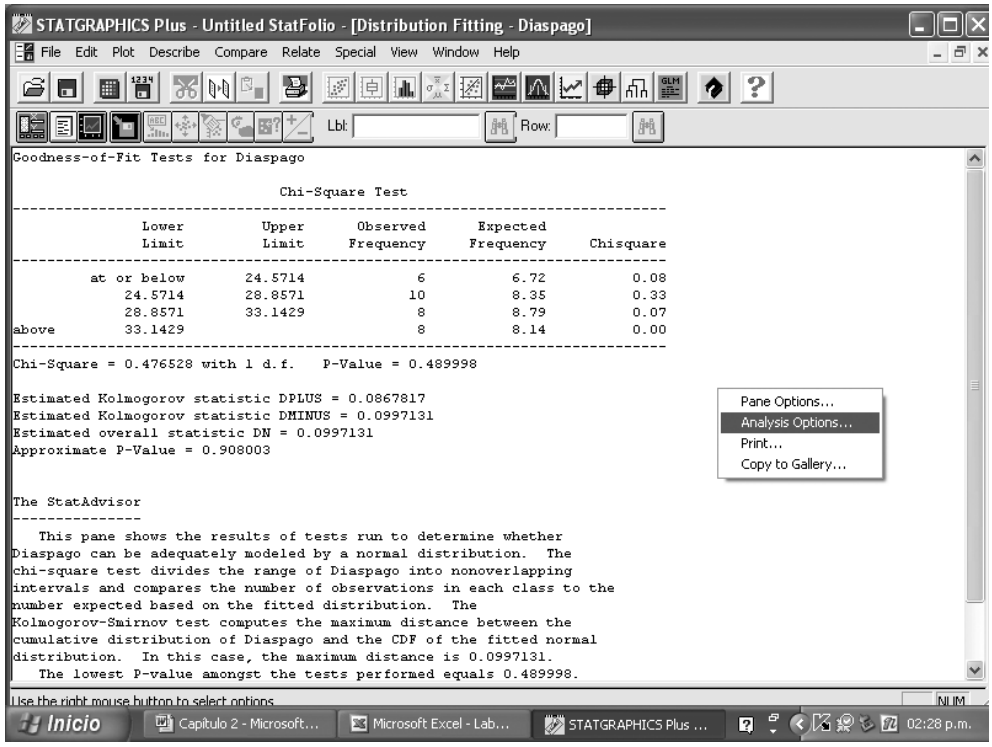

STATGRAPHICS Plus - Untitled StatFolio - [<untitled>]
File Edit Plot Describe Compare Relate Special View Window Help
Numeric Data
Categorical Data
Hypothesis Tests...
Sample Size Determination...
One-Variable Analysis...
Multiple-Variable Analysis...
Distribution Fitting...
Col_5 Col_6 Col_7
Distribution fitting for numeric data
NUM
Inicio
Capítulo 2 - Microsoft...
Microsoft Excel - Lab...
STATGRAPHICS Plus ...
02:15 p.m.

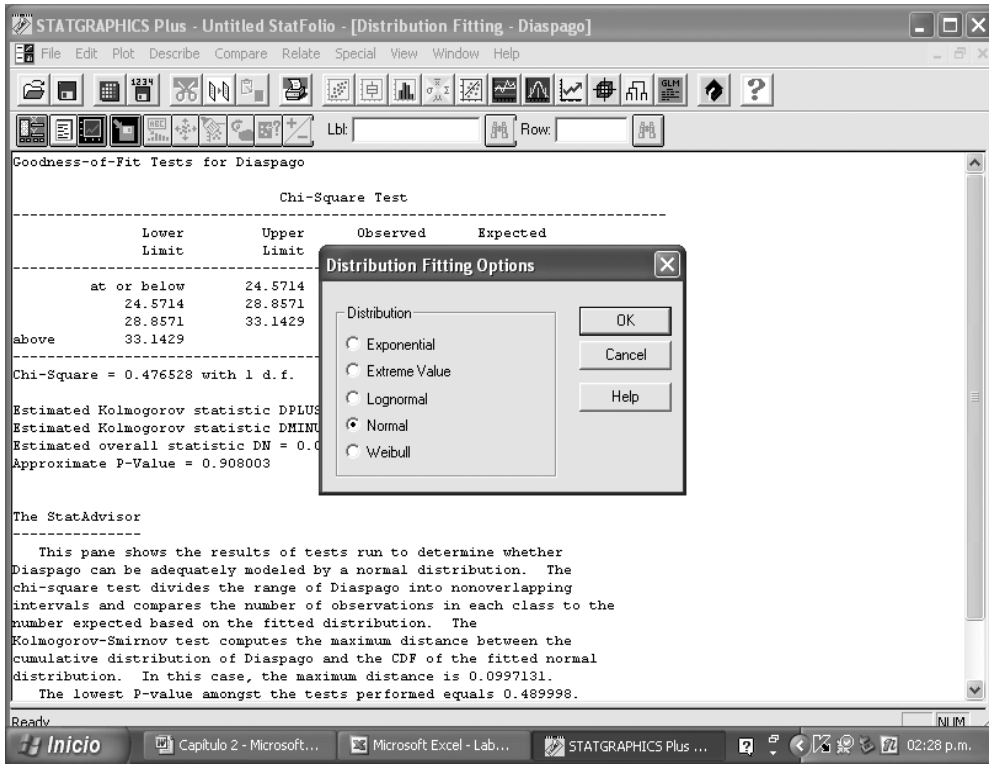

STATGRAPHICS Plus - Untitled StatFolio - [<untitled>]
File Edit Plot Describe Compare Relate Special View Window Help
Diaspago Col_2 Col_3 Col_4 Col_5 Col_6 Col_7
1 35
2 32
3 28
4 36
5 34
6 23
7 26
8 41
9 35
10 37
11 42
12 24
13 19
14 34
15 18
16 32
17 27
18 26
19 29
20 31
21 28
22 18
Distribution Fitting
Diaspago
Data:
Diaspago
(Select:)
Sort
OK
Cancel
Delete
Transform...
Help
Ready
NUM
Inicio
Capítulo 2 - Microsoft...
Microsoft Excel - Lab...
STATGRAPHICS Plus ...
02:16 p.m.

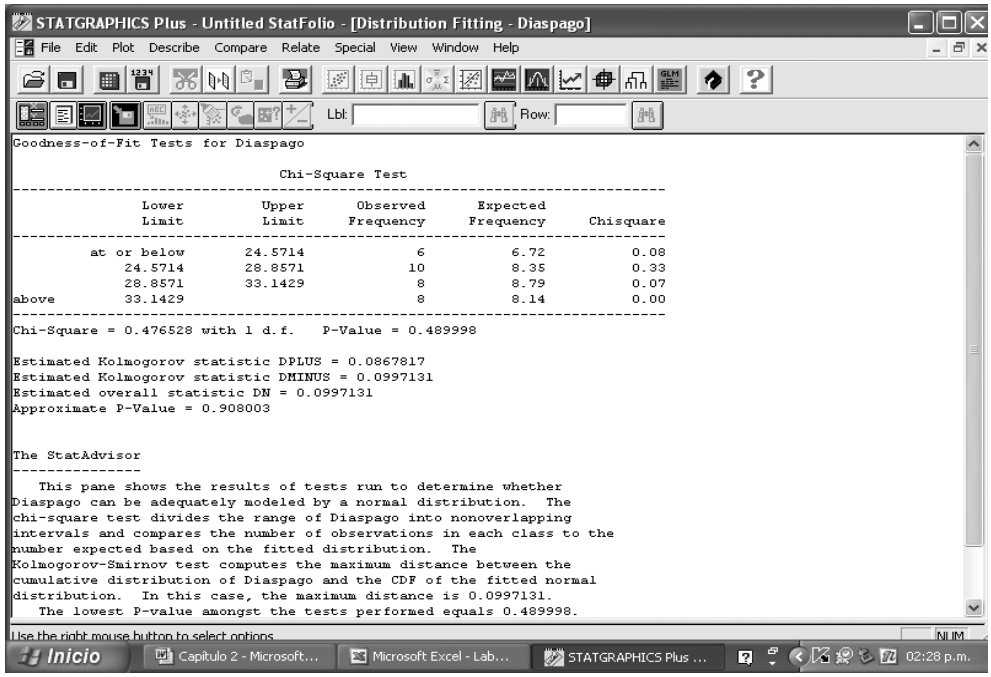

STATGRAPHICS Plus - Untitled StatFolio - [Distribution Fitting - Diaspago]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl:
Row:
Analysis
Tabular options
Data variable: Diaspago
32 values ranging from 18.0 to 42.0
Fitted normal distribution:
mean = 29.2813
standard deviation = 5.8433
The StatAdvisor
This analysis shows the results of fitting a normal distribution to the data on Diaspago. The estimated parameters of the fitted distribution are shown above. You can test whether the normal distribution fits the data adequately by selecting Goodness-of-Fit Tests from the list of Tabular Options. You can also assess visually how well the normal distribution fits by selecting Frequency Histogram from the list of Graphical Options. Other options within the procedure allow you to compute and display tail areas and critical values for the distribution. To select a different distribution, press the alternate mouse button and select Analysis Options.
Tabular Options
NUM
Inicio
Capítulo 2 - Microsoft...
Microsoft Excel - Lab...
STATGRAPHICS Plus ...
02:16 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [Distribution Fitting - Diaspago]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Analysis Summary
Data variable: Diaspago
32 values ranging from 18.0 to 42.9
Fitted normal distribution:
mean = 29.2813
standard deviation = 5.8433
The StatAdvisor
This analysis shows the result
the data on Diaspago. The estim
distribution are shown above. Yo
distribution fits the data adequ
Tests from the list of Tabular Op
how well the normal distribution
from the list of Graphical Optio
procedure allow you to compute and display tail areas and critical
values for the distribution. To select a different distribution,
press the alternate mouse button and select Analysis Options.
Tabular Options
Analysis Summary
Tests for Normality
Goodness-of-Fit Tests
Tail Areas
Critical Values
OK
Cancel
All
Help
Use the right mouse button to select options
NUM
Inicio
Capítulo 2 - Microsoft...
Microsoft Excel - Lab...
STATGRAPHICS Plus ...
02:16 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [Distribution Fitting - Diaspago]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Goodness-of-Fit Tests for Diaspago
Chi-Square Test
Lower Limit | Upper Limit | Observed Frequency | Expected Frequency | Chisquare
at or below 24.5714 6 6.72 0.08
24.5714 28.8571 10 8.35 0.33
28.8571 33.1429 8 8.79 0.07
above 33.1429 8 8.14 0.00
Chi-Square = 0.476528 with 1 d.f. P-Value = 0.489998
Estimated Kolmogorov statistic DPLUS = 0.0867817
Estimated Kolmogorov statistic DMINUS = 0.0997131
Estimated overall statistic DN = 0.0997131
Approximate P-Value = 0.908003
The StatAdvisor
This pane shows the results of tests run to determine whether
Diaspago can be adequately modeled by a normal distribution. The
chi-square test divides the range of Diaspago into nonoverlapping
intervals and compares the number of observations in each class to the
number expected based on the fitted distribution. The
Kolmogorov-Smirnov test computes the maximum distance between the
cumulative distribution of Diaspago and the CDF of the fitted normal
distribution. In this case, the maximum distance is 0.0997131.
The lowest P-value amongst the tests performed equals 0.489998.
Use the right mouse button to select options
NUM
Inicio
Capítulo 2 - Microsoft...
Microsoft Excel - Lab...
STATGRAPHICS Plus ...
02:28 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [Distribution Fitting - Diaspago]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Goodness-of-Fit Tests for Diaspago
Chi-Square Test
Lower Limit Upper Limit Observed Frequency Expected Frequency Chisquare
at or below 24.5714 6 6.72 0.08
24.5714 28.8571 10 8.35 0.33
28.8571 33.1429 8 8.79 0.07
above 33.1429 8 8.14 0.00
Chi-Square = 0.476528 with 1 d.f. P-Value = 0.489998
Estimated Kolmogorov statistic DPLUS = 0.0867817
Estimated Kolmogorov statistic DMINUS = 0.0997131
Estimated overall statistic DN = 0.0997131
Approximate P-Value = 0.908003
Pane Options...
Analysis Options...
Print...
Copy to Gallery...
The StatAdvisor
This pane shows the results of tests run to determine whether Diaspago can be adequately modeled by a normal distribution. The chi-square test divides the range of Diaspago into nonoverlapping intervals and compares the number of observations in each class to the number expected based on the fitted distribution. The Kolmogorov-Smirnov test computes the maximum distance between the cumulative distribution of Diaspago and the CDF of the fitted normal distribution. In this case, the maximum distance is 0.0997131.
The lowest P-value amongst the tests performed equals 0.489998.
Use the right mouse button to select options
Inicio
Capítulo 2 - Microsoft...
Microsoft Excel - Lab...
STATGRAPHICS Plus ...
02:28 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [Distribution Fitting - Diaspago]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Goodness-of-Fit Tests for Diaspago
Chi-Square Test
Lower Limit Upper Limit Observed Expected
at or below 24.5714
24.5714 28.8571
28.8571 33.1429
above 33.1429
Chi-Square = 0.476528 with 1 d.f.
Approximate P-Value = 0.908003
Distribution Fitting Options
Distribution
Exponential
Extreme Value
Lognormal
Normal
Weibull
OK
Cancel
Help
The StatAdvisor
This pane shows the results of tests run to determine whether Diaspago can be adequately modeled by a normal distribution. The chi-square test divides the range of Diaspago into nonoverlapping intervals and compares the number of observations in each class to the number expected based on the fitted distribution. The Kolmogorov-Smirnov test computes the maximum distance between the cumulative distribution of Diaspago and the CDF of the fitted normal distribution. In this case, the maximum distance is 0.0997131.
The lowest P-value amongst the tests performed equals 0.489998.
Ready
Inicio
Capítulo 2 - Microsoft...
Microsoft Excel - Lab...
STATGRAPHICS Plus ...
02:28 p.m.



Véase que ante todo, se cumple el requisito para realizar esta dócima, pues como la cantidad de clases en la tabla de distribución de frecuencias es menor que 5 (se observan 4 intervalos de clases) entonces todas las frecuencias esperadas deben tener un valor igual o mayor que 5, tal y como se muestra (6.72, 8.35, 8.79 y 8.14). El valor de probabilidad de la prueba Chi-Cuadrado para la Bondad de Ajuste, es igual a 0.49. Como este valor no es menor que 0.01, entonces no se cumple la región crítica y se acepta la hipótesis nula. El jefe del Departamento de Riesgo puede afirmar finalmente que, la cantidad de días de demora de pago de los TT.OO., sigue una distribución normal con un nivel de significación del 1%.

5.4. Prueba de Kolmogorov-Smirnov.

Esta es una prueba muy similar a la X^2 de Pearson demostrada en el ejemplo anterior.

Véase un ejemplo.

Ejemplo 2:

En la Agencia de Viajes B ubicada en el polo turístico de Cayo Coco, el

Departamento de Opcionales desea saber si las edades de los turistas que compraron la excursión "Trinidad" el mes anterior, siguen o no una distribución normal.

Para ello, los integrantes del departamento recogieron la edad de cada uno de los turistas que diariamente fueron a Trinidad en los últimos 31 días, calculando un promedio de edad diario que viabilice el análisis de los datos, con un nivel de seguridad del 95%. Los datos se hallan a continuación:

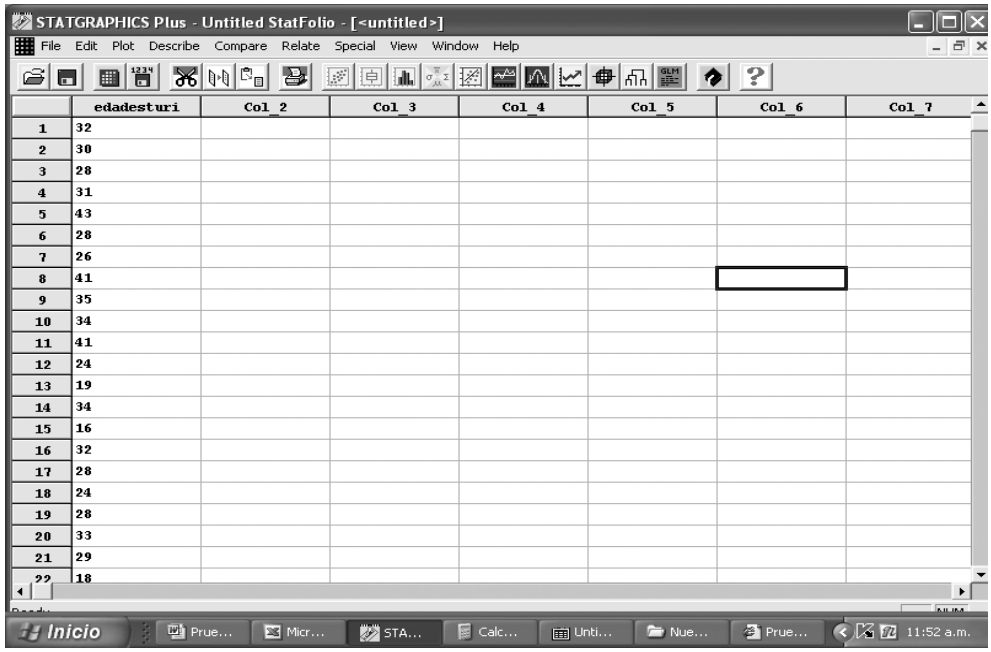
Días del mes	Promedio de edades
1	32
2	30
3	28
4	31
5	43
6	28
7	26
8	41
9	35
10	34
11	41
12	24
13	19
14	34
15	16
16	32
17	28
18	24
19	28
20	33
21	29
22	18
23	24
24	31
25	30
26	22
27	27
28	26
29	30
30	28
31	25

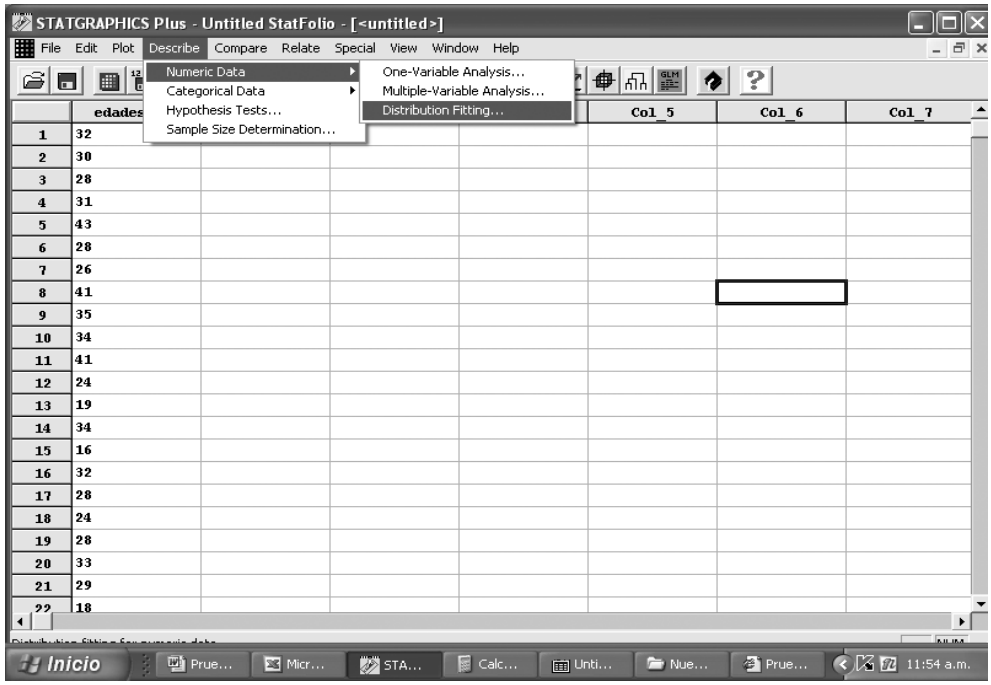
Solución:

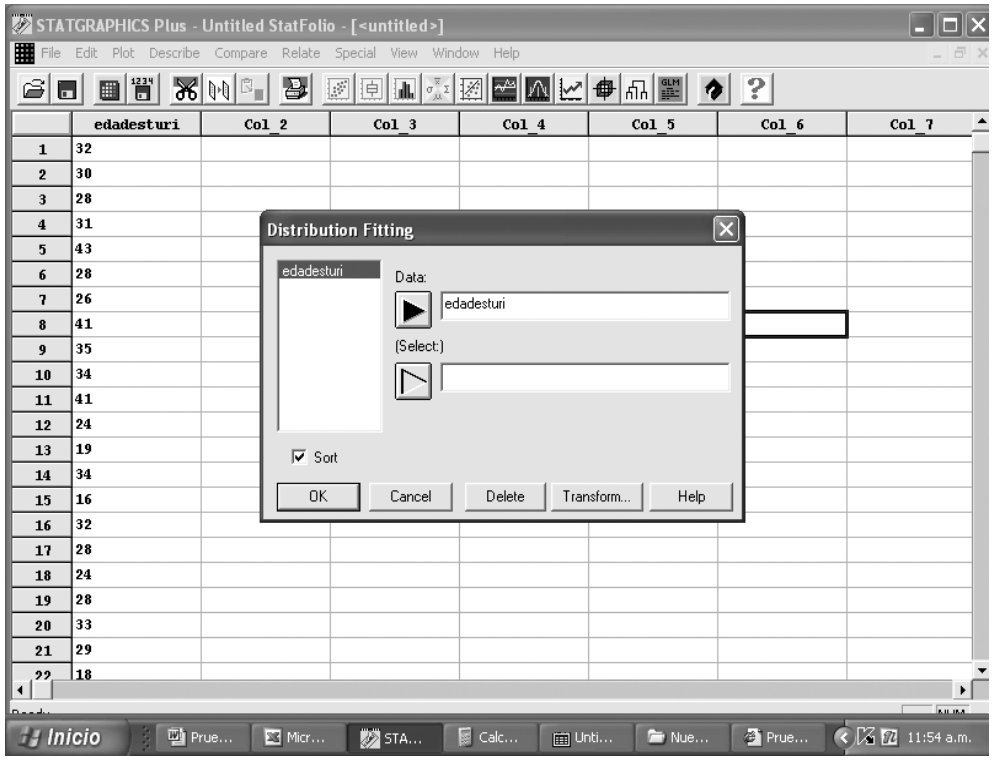
Variable de estudio X: edad de los turistas.

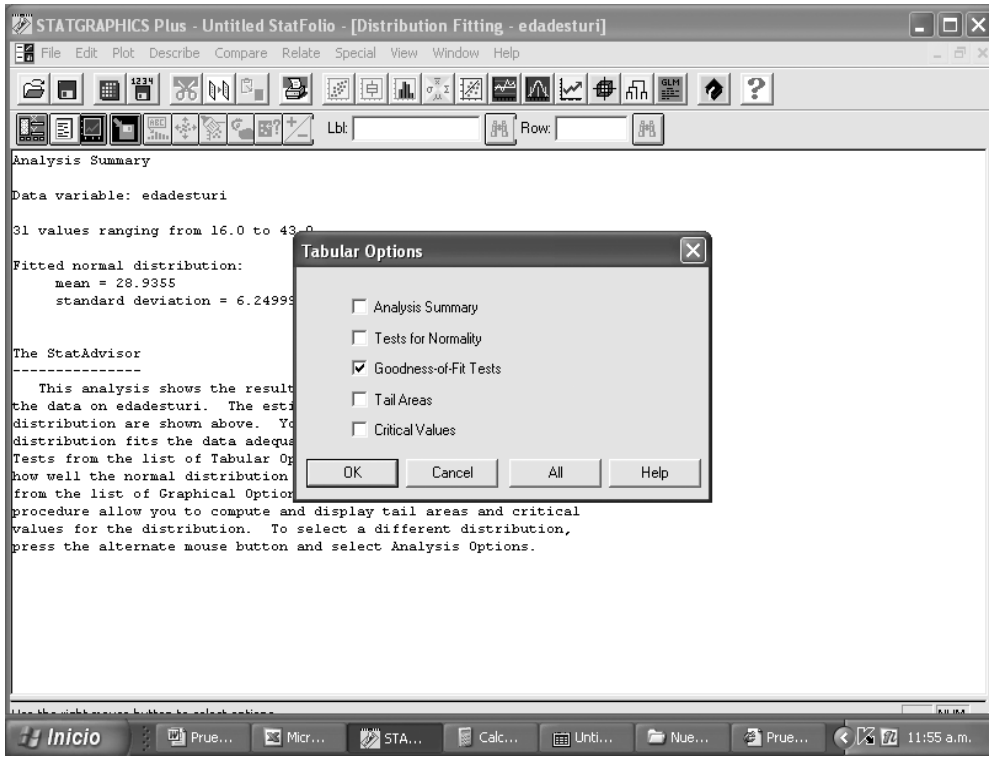
H_0: X ~ N (μ; σ^2) (la edad sigue una distribución normal)

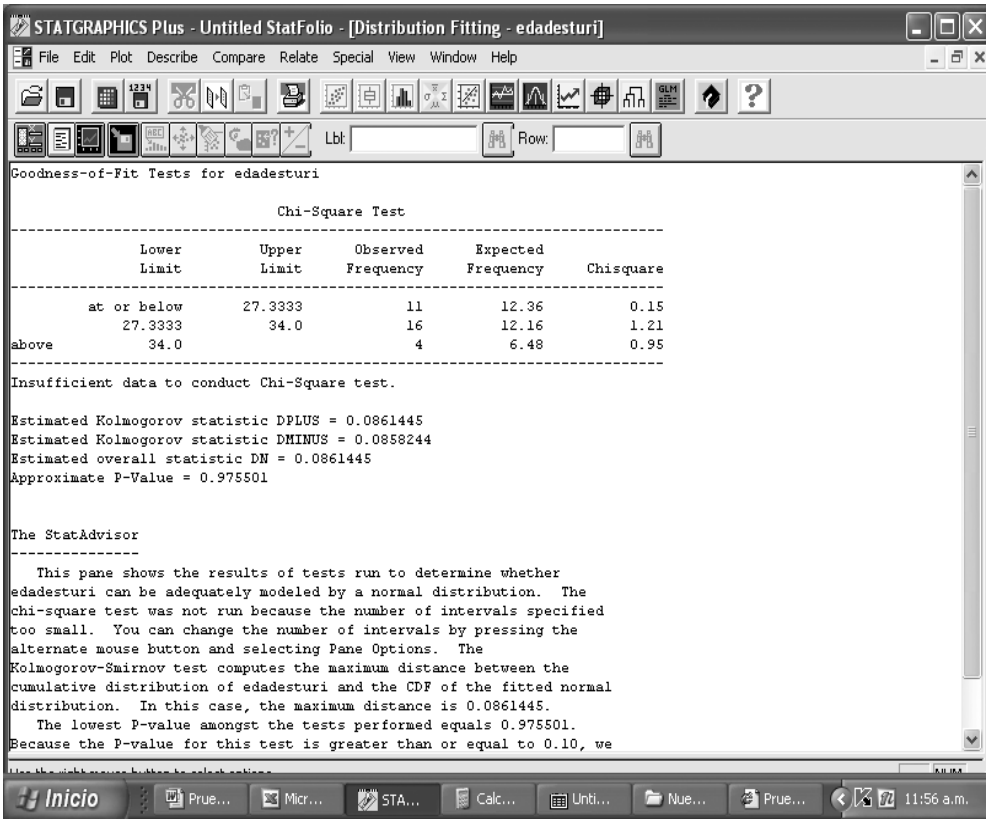
H_1: X no~ N (μ; σ^2) (la edad no sigue una distribución normal)
Utilizando el STATGRAPHICS Plus para realizar esta dócima de hipótesis no paramétrica, sería:





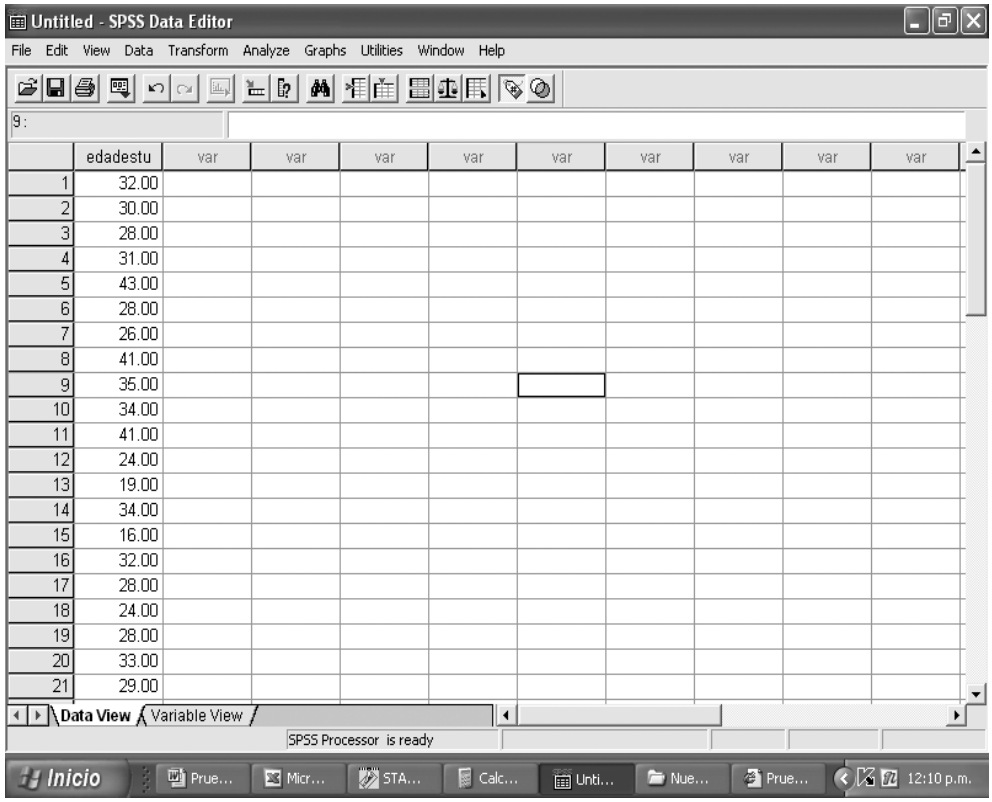

STATGRAPHICS Plus - Untitled StatFolio - [<untitled>]
File Edit Plot Describe Compare Relate Special View Window Help
edadesturi Col_2 Col_3 Col_4 Col_5 Col_6 Col_7
1 32
2 30
3 28
4 31
5 43
6 28
7 26
8 41
9 35
10 34
11 41
12 24
13 19
14 34
15 16
16 32
17 28
18 24
19 28
20 33
21 29
22 18
Distribution Fitting
edadesturi
Data:
edadesturi
(Select:)
Sort
OK Cancel Delete Transform... Help
Inicio
11:54 a.m.

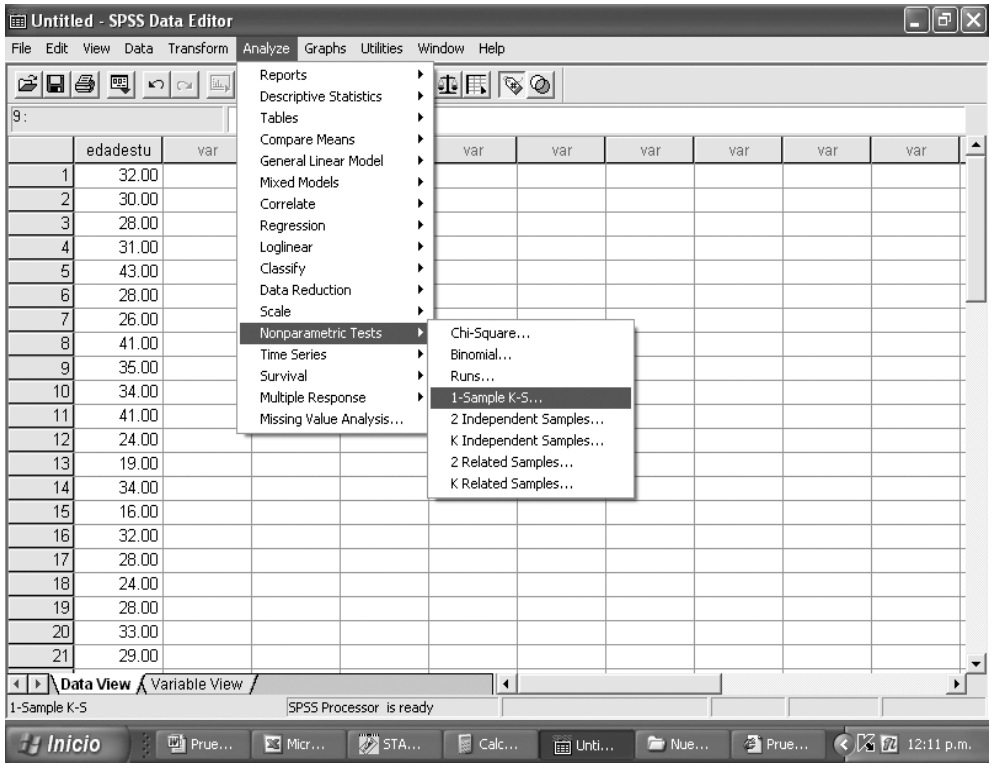

STATGRAPHICS Plus - Untitled StatFolio - [Distribution Fitting - edadesturi]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Analysis Summary
Data variable: edadesturi
31 values ranging from 16.0 to 43.0
Fitted normal distribution:
mean = 28.9355
standard deviation = 6.24995
The StatAdvisor
This analysis shows the result
the data on edadesturi. The esti
distribution are shown above. Yo
distribution fits the data adequa
Tests from the list of Tabular Op
how well the normal distribution
from the list of Graphical Optio
procedure allow you to compute and display tail areas and critical
values for the distribution. To select a different distribution,
press the alternate mouse button and select Analysis Options.
Tabular Options
Analysis Summary
Tests for Normality
Goodness-of-Fit Tests
Tail Areas
Critical Values
OK Cancel All Help
Inicio
11:55 a.m.

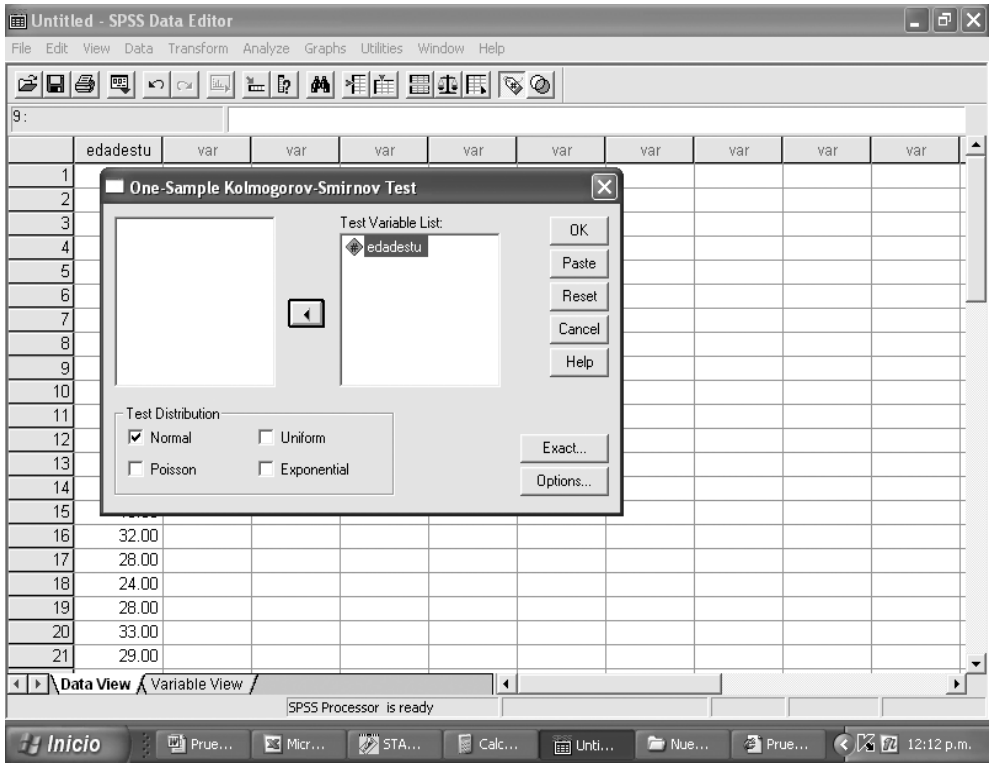


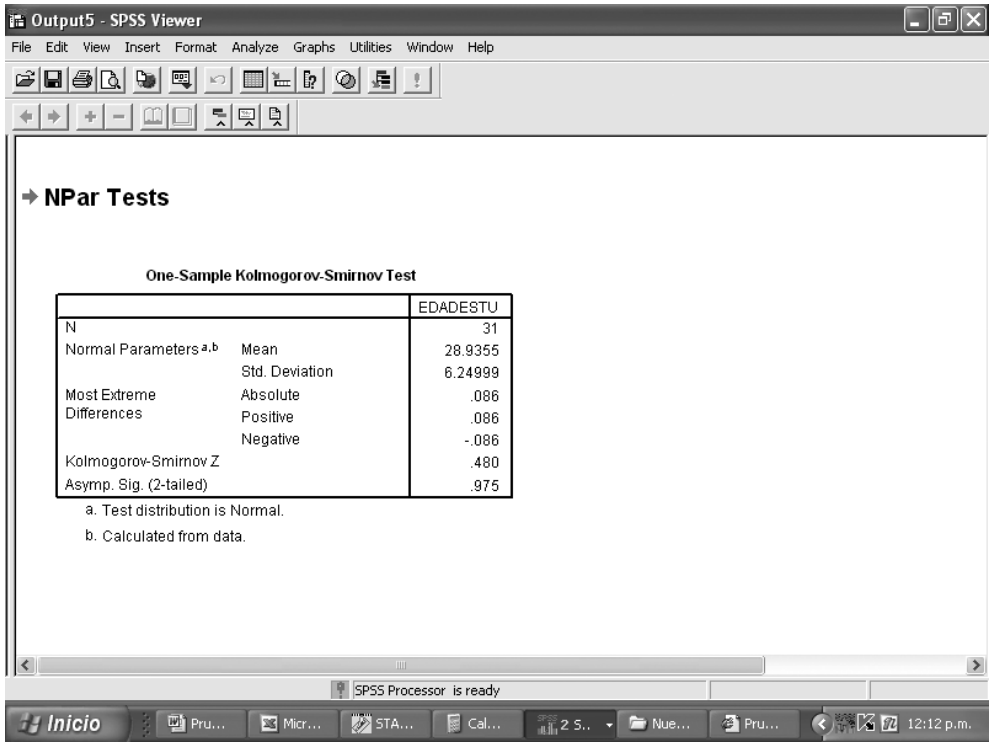
Según se observa, el valor de probabilidad de la dócima de Kolmogorov-Smirnov es igual a 0.98. Como ese valor de probabilidad no es menor que 0.05, entonces no se cumple la región crítica y se acepta la hipótesis nula. Finalmente, los integrantes del Departamento de Opcionales de la agencia, pueden afirmar que la excursión "Trinidad" es comprada por turistas de edades similares, lo cual evidencia que las mismas siguen una distribución normal con un nivel de significación del 5%.

Ahora obsérvese esta misma prueba realizada con el SPSS:


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
9 :
edadestu
var
1 32.00
2 30.00
3 28.00
4 31.00
5 43.00
6 28.00
7 26.00
8 41.00
9 35.00
10 34.00
11 41.00
12 24.00
13 19.00
14 34.00
15 16.00
16 32.00
17 28.00
18 24.00
19 28.00
20 33.00
21 29.00
Data View
Variable View
SPSS Processor is ready
Inicio
12:10 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Chi-Square...
Binomial...
Runs...
1-Sample K-S...
2 Independent Samples...
K Independent Samples...
2 Related Samples...
K Related Samples...
edadestu
Data View
Variable View
1-Sample K-S
SPSS Processor is ready
Inicio
12:11 p.m.


Untitled - SPSS Data Editor
One-Sample Kolmogorov-Smirnov Test
Test Variable List:
edadestu
OK
Paste
Reset
Cancel
Help
Test Distribution
Normal
Uniform
Poisson
Exponential
Exact...
Options...
Data View
Variable View
SPSS Processor is ready
Inicio
12:12 p.m.



5.5. Prueba de Shapiro-Wilk.

Véase un ejemplo.

Ejemplo 3:

El Restaurante Y perteneciente a la Cadena Palmares ubicado en la zona turística de La Habana Vieja, ha obtenido en los últimos meses, mejores resultados en cuanto a sus niveles de utilidades.

La dirección del restaurante ha comprobado que los niveles de ganancia han mostrado un ascenso considerable, y que en los últimos 20 días del mes en curso, se han mantenido, sospechando que dichos niveles de utilidades han seguido una distribución normal con un nivel de significación del 1%. En la siguiente tabla se observan los datos recopilados:

Días del mes	Utilidades en CUC
1	230.60
2	335.05
3	117.35
4	265.90
5	189.45
6	245.00
7	312.20
8	165.75
9	284.90
10	132.10
11	279.65
12	234.85
13	322.25
14	157.65
15	278.35
16	123.35
17	345.95
18	132.65
19	299.85
20	188.15

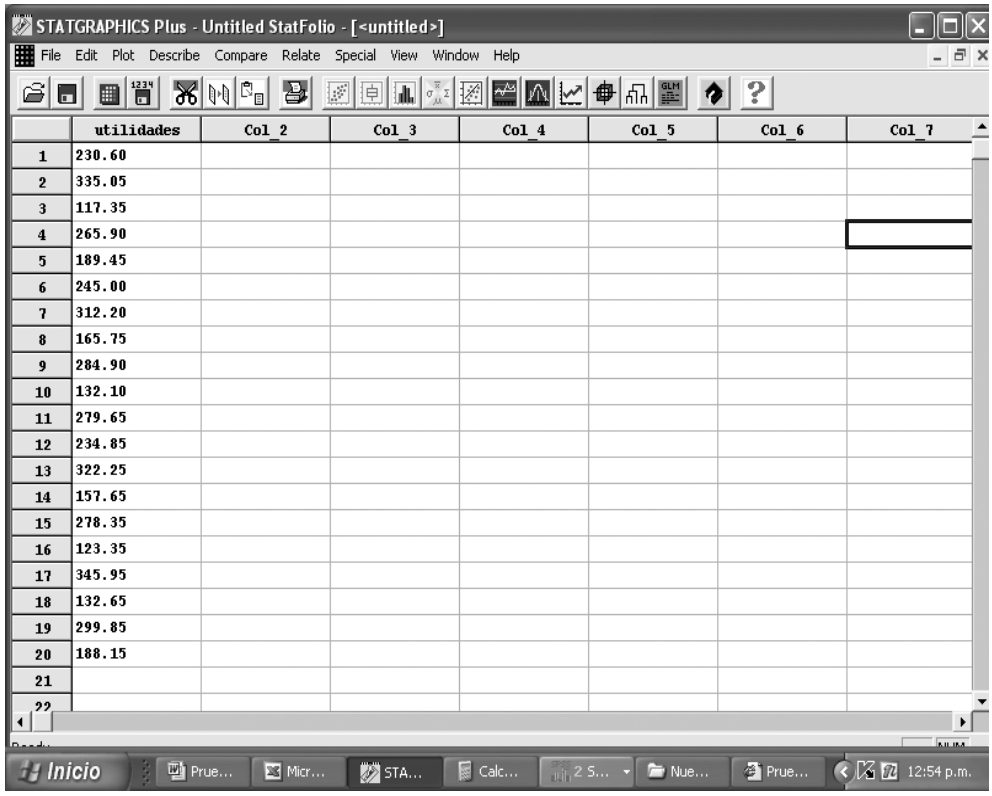
Solución:

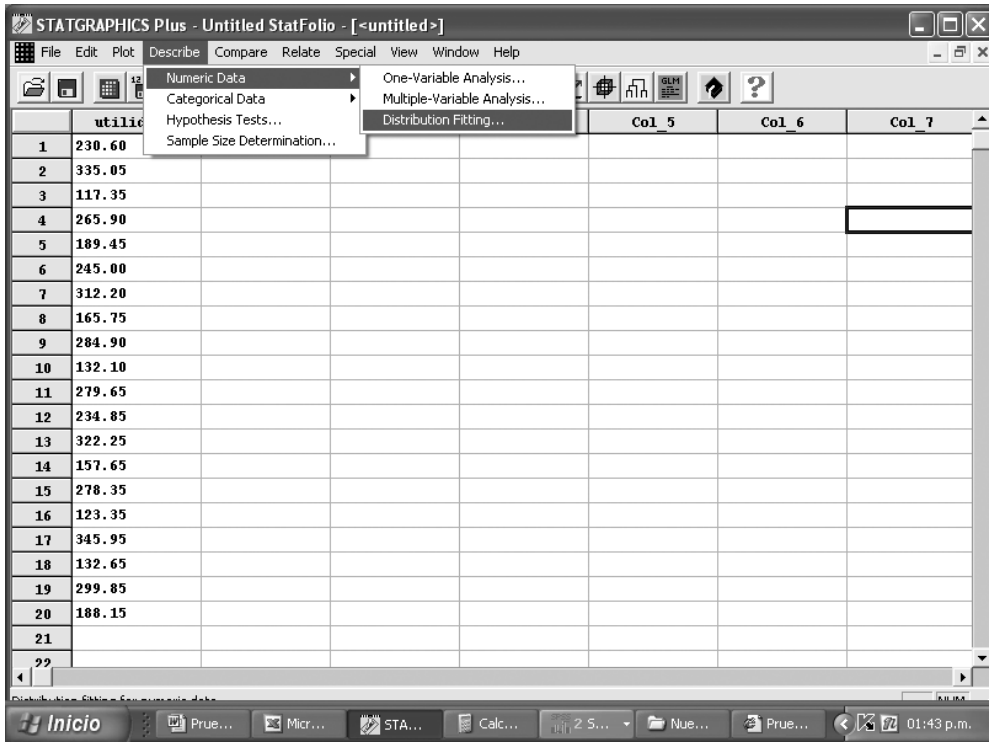
Variable de estudio X: nivel de utilidades en CUC.

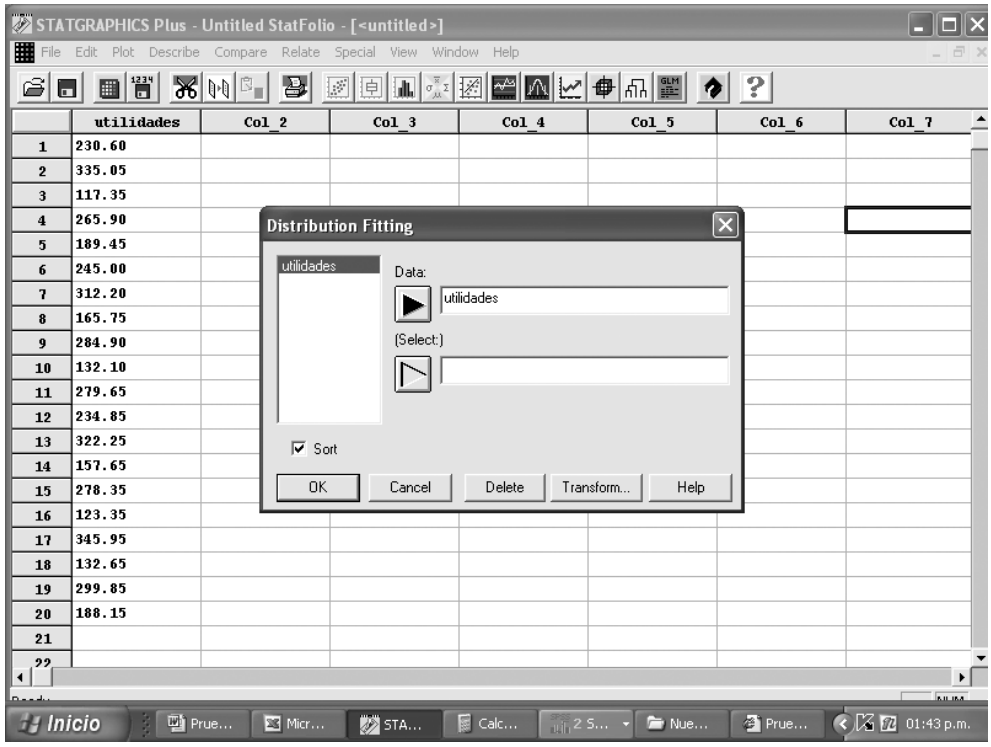
H_0: $X \sim N(\mu; \sigma^2)$ (el nivel de utilidades sigue una distribución normal)

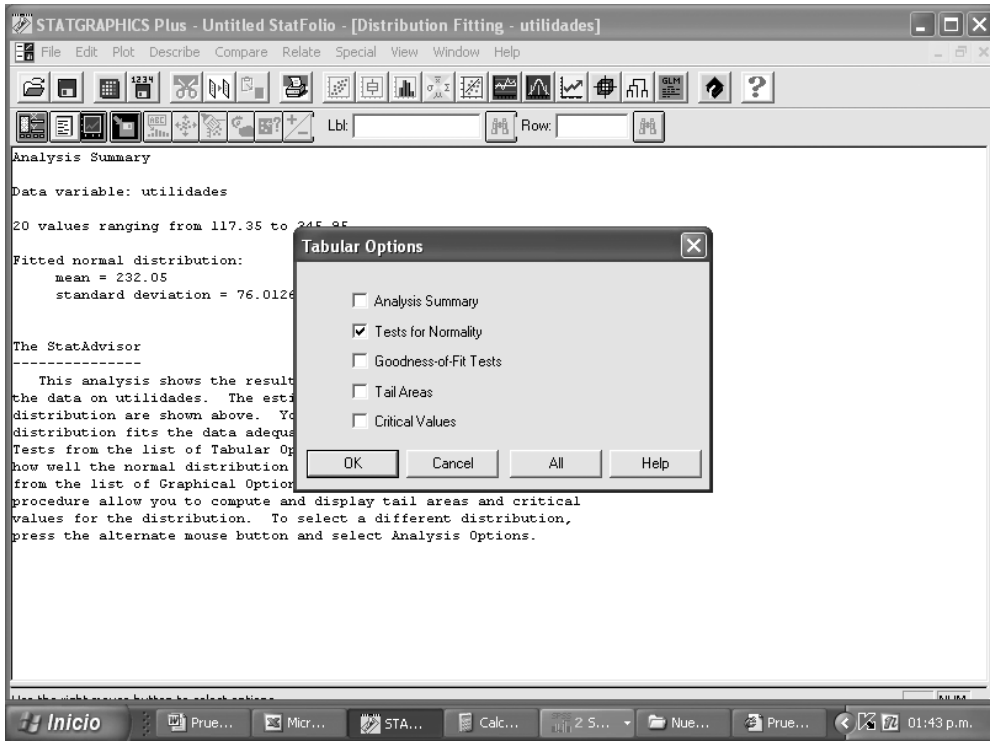
H_1: $X \overset{no}{\sim} N(\mu; \sigma^2)$ (el nivel de utilidades no sigue una distribución normal)

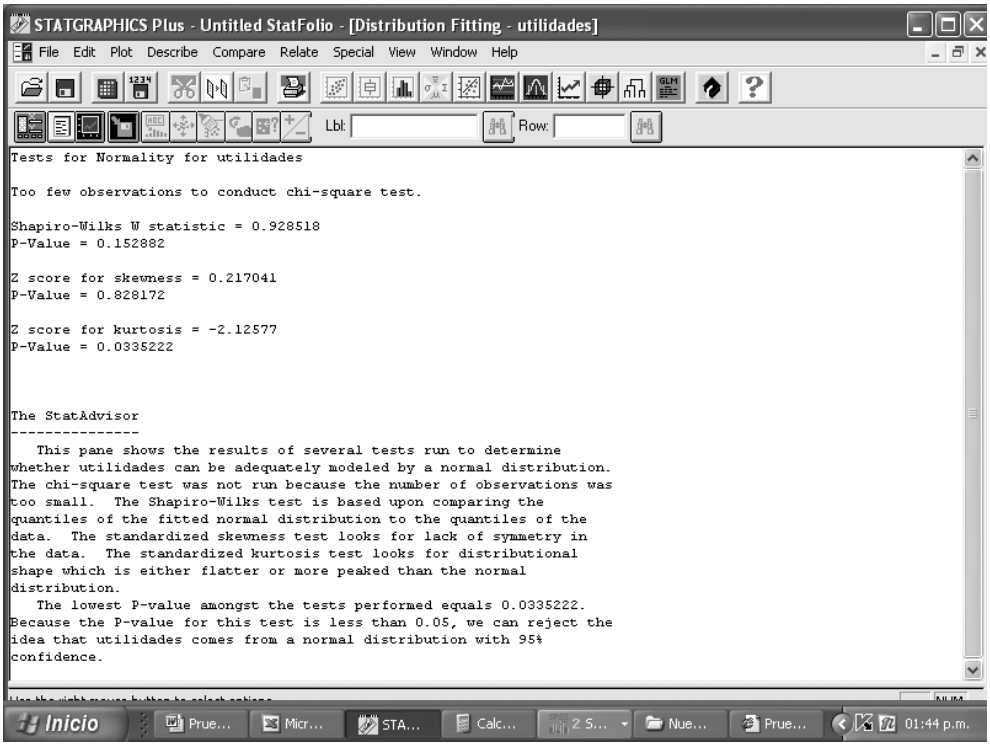
Utilizando el STATGRAPHICS Plus para realizar esta dócima de hipótesis no paramétrica, sería:


STATGRAPHICS Plus - Untitled StatFolio - [<untitled>]
File Edit Plot Describe Compare Relate Special View Window Help
utilidades Col_2 Col_3 Col_4 Col_5 Col_6 Col_7
1 230.60
2 335.05
3 117.35
4 265.90
5 189.45
6 245.00
7 312.20
8 165.75
9 284.90
10 132.10
11 279.65
12 234.85
13 322.25
14 157.65
15 278.35
16 123.35
17 345.95
18 132.65
19 299.85
20 188.15
21
Inicio
12:54 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [<untitled>]
File Edit Plot Describe Compare Relate Special View Window Help
Numeric Data
Categorical Data
Hypothesis Tests...
Sample Size Determination...
One-Variable Analysis...
Multiple-Variable Analysis...
Distribution Fitting...
Col_5 Col_6 Col_7
1 230.60
2 335.05
3 117.35
4 265.90
5 189.45
6 245.00
7 312.20
8 165.75
9 284.90
10 132.10
11 279.65
12 234.85
13 322.25
14 157.65
15 278.35
16 123.35
17 345.95
18 132.65
19 299.85
20 188.15
21
Inicio
01:43 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [<untitled>]
File Edit Plot Describe Compare Relate Special View Window Help
utilidades Col_2 Col_3 Col_4 Col_5 Col_6 Col_7
1 230.60
2 335.05
3 117.35
4 265.90
5 189.45
6 245.00
7 312.20
8 165.75
9 284.90
10 132.10
11 279.65
12 234.85
13 322.25
14 157.65
15 278.35
16 123.35
17 345.95
18 132.65
19 299.85
20 188.15
21
Distribution Fitting
utilidades
Data:
utilidades
(Select:)
Sort
OK Cancel Delete Transform... Help
Inicio
01:43 p.m.

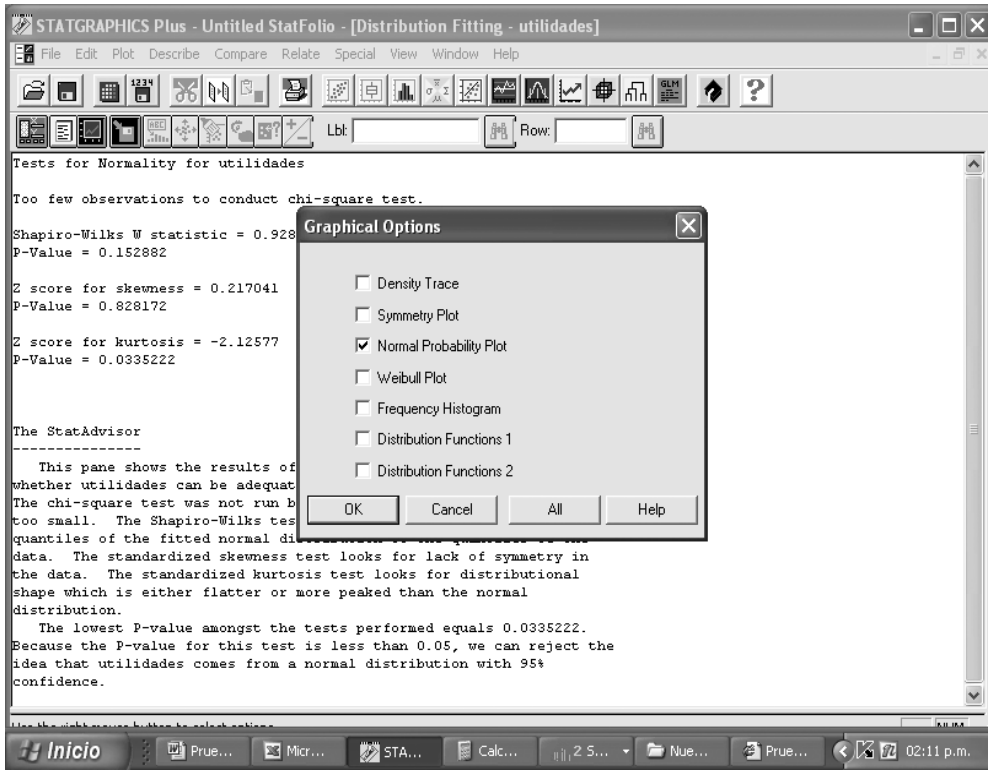

STATGRAPHICS Plus - Untitled StatFolio - [Distribution Fitting - utilidades]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Analysis Summary
Data variable: utilidades
20 values ranging from 117.35 to 345.95
Fitted normal distribution:
mean = 232.05
standard deviation = 76.0126
The StatAdvisor
This analysis shows the result
the data on utilidades. The esti
distribution are shown above. Yo
distribution fits the data adequ
Tests from the list of Tabular Op
how well the normal distribution
from the list of Graphical Optio
procedure allow you to compute and display tail areas and critical
values for the distribution. To select a different distribution,
press the alternate mouse button and select Analysis Options.
Tabular Options
Analysis Summary
Tests for Normality
Goodness-of-Fit Tests
Tail Areas
Critical Values
OK Cancel All Help
Inicio
01:43 p.m.

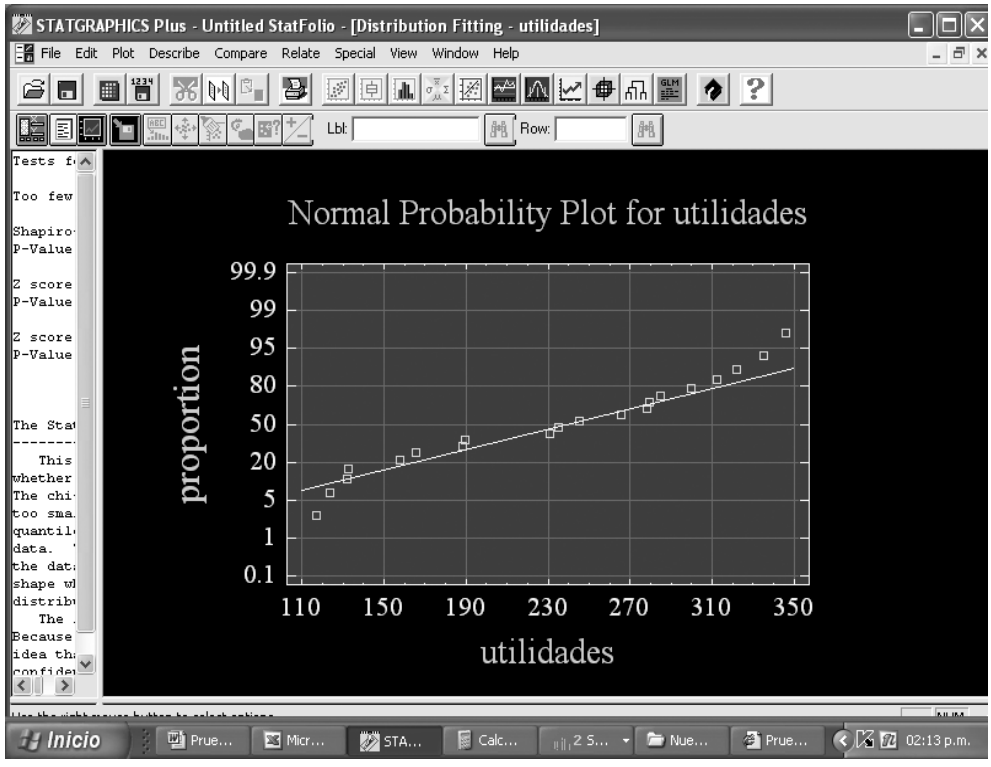


Según se observa, el valor de probabilidad de la dócima de Shapiro-Wilk es igual a 0.15. Como ese valor de probabilidad no es menor que 0.01, entonces no se cumple la región crítica y se acepta la hipótesis nula.

Puede entonces la dirección del restaurante afirmar que, efectivamente, los niveles mantenidos de ascenso en las ganancias, siguen una distribución normal con un nivel de significación del 1%.

También puede observarse el gráfico que acompaña esta dócima, y que permite apreciar el ajuste a dicha distribución normal en este ejemplo.

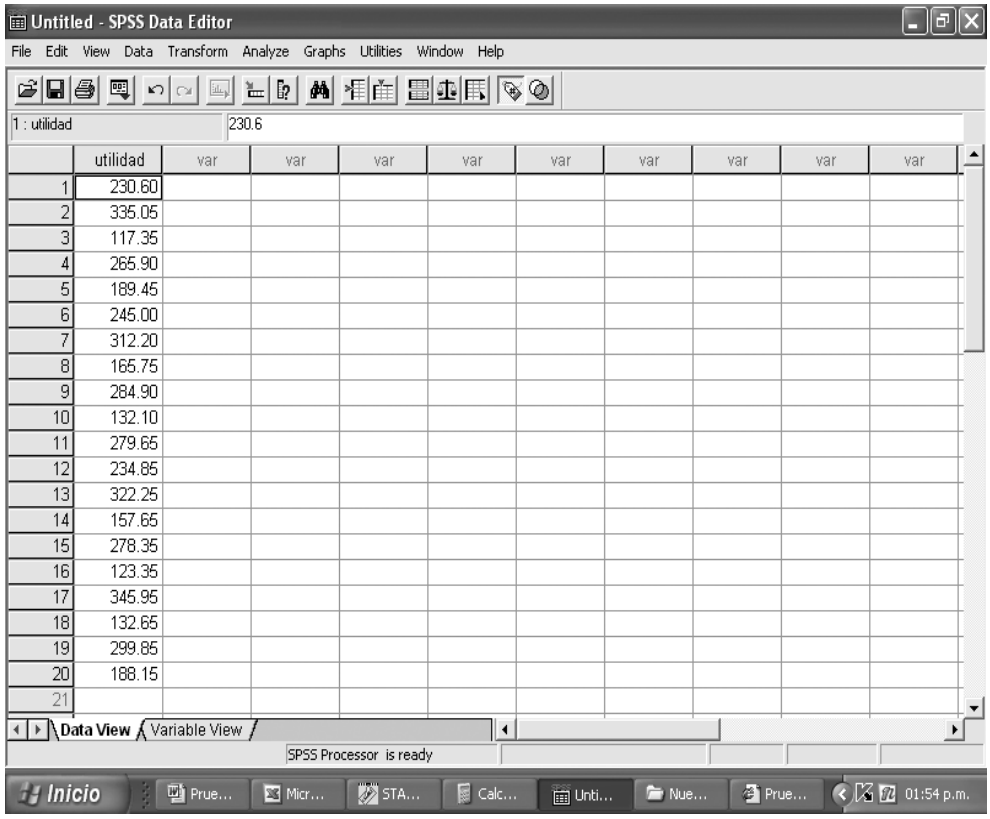

STATGRAPHICS Plus - Untitled StatFolio - [Distribution Fitting - utilidades]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Tests for Normality for utilidades
Too few observations to conduct chi-square test.
Shapiro-Wilks W statistic = 0.928
P-Value = 0.152882
Z score for skewness = 0.217041
P-Value = 0.828172
Z score for kurtosis = -2.12577
P-Value = 0.0335222
The StatAdvisor
This pane shows the results of
whether utilidades can be adequat
The chi-square test was not run b
too small. The Shapiro-Wilks tes
quantiles of the fitted normal di
data. The standardized skewness test looks for lack of symmetry in
the data. The standardized kurtosis test looks for distributional
shape which is either flatter or more peaked than the normal
distribution.
The lowest P-value amongst the tests performed equals 0.0335222.
Because the P-value for this test is less than 0.05, we can reject the
idea that utilidades comes from a normal distribution with 95%
confidence.
Graphical Options
Density Trace
Symmetry Plot
Normal Probability Plot
Weibull Plot
Frequency Histogram
Distribution Functions 1
Distribution Functions 2
OK
Cancel
All
Help
Inicio
02:11 p.m.

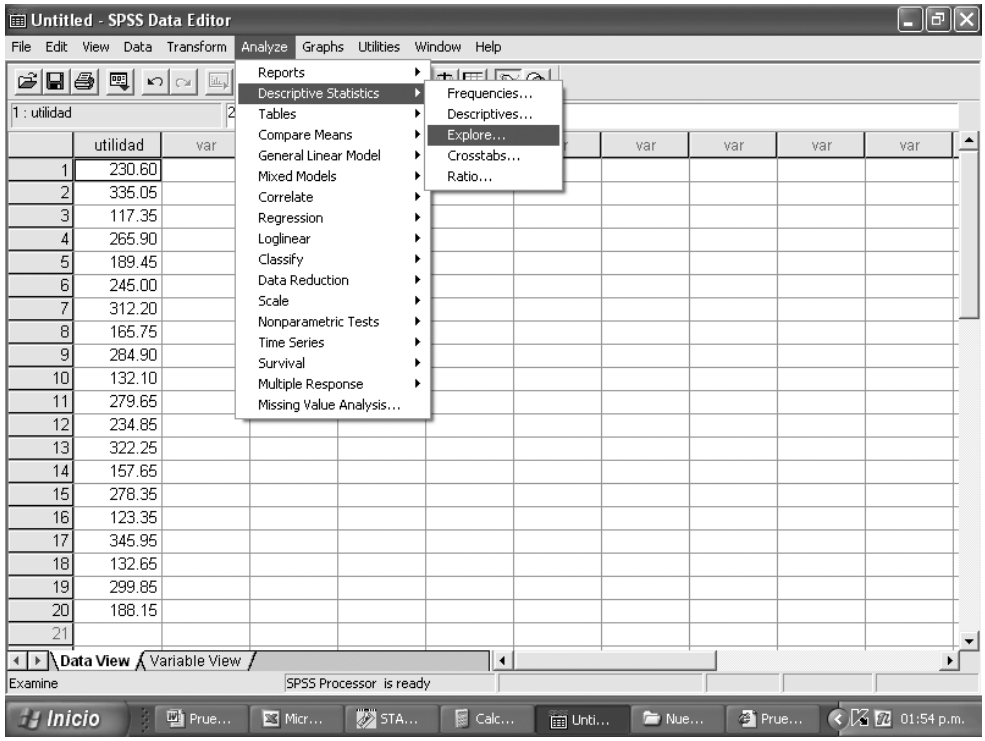

STATGRAPHICS Plus - Untitled StatFolio - [Distribution Fitting - utilidades]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Normal Probability Plot for utilidades
proportion
99.9
99
95
80
50
20
5
1
0.1
110
150
190
230
270
310
350
utilidades
Inicio
02:13 p.m.

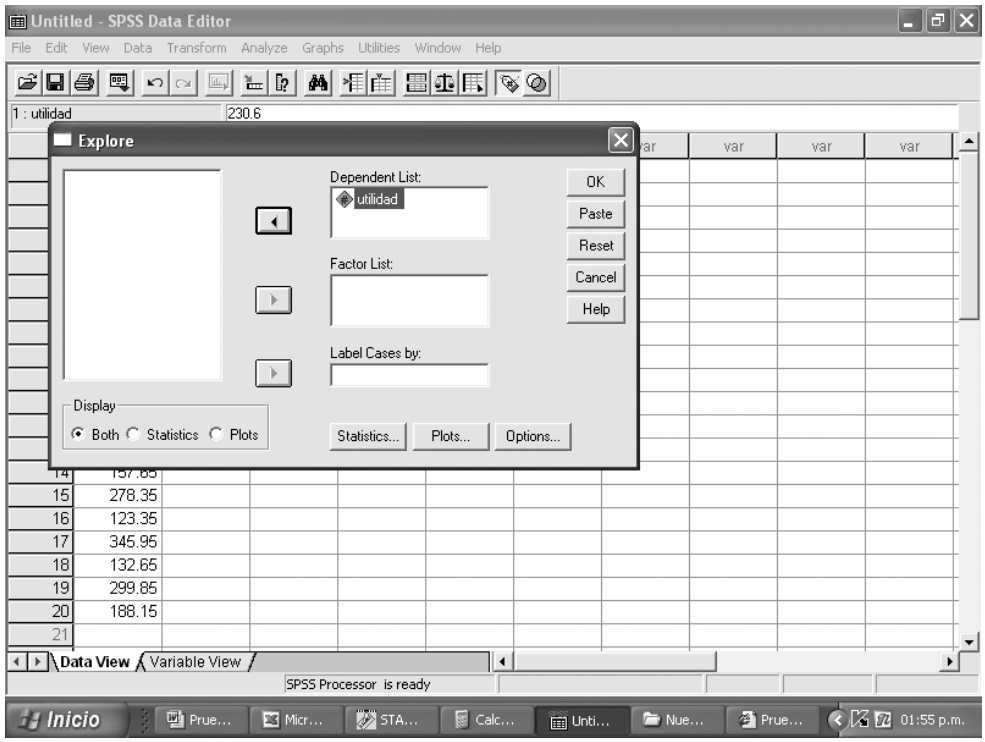
Véase que los pequeños cuadraditos que representan las utilidades promedio diarias durante 20 días, están casi alineadas con la recta que se observa.

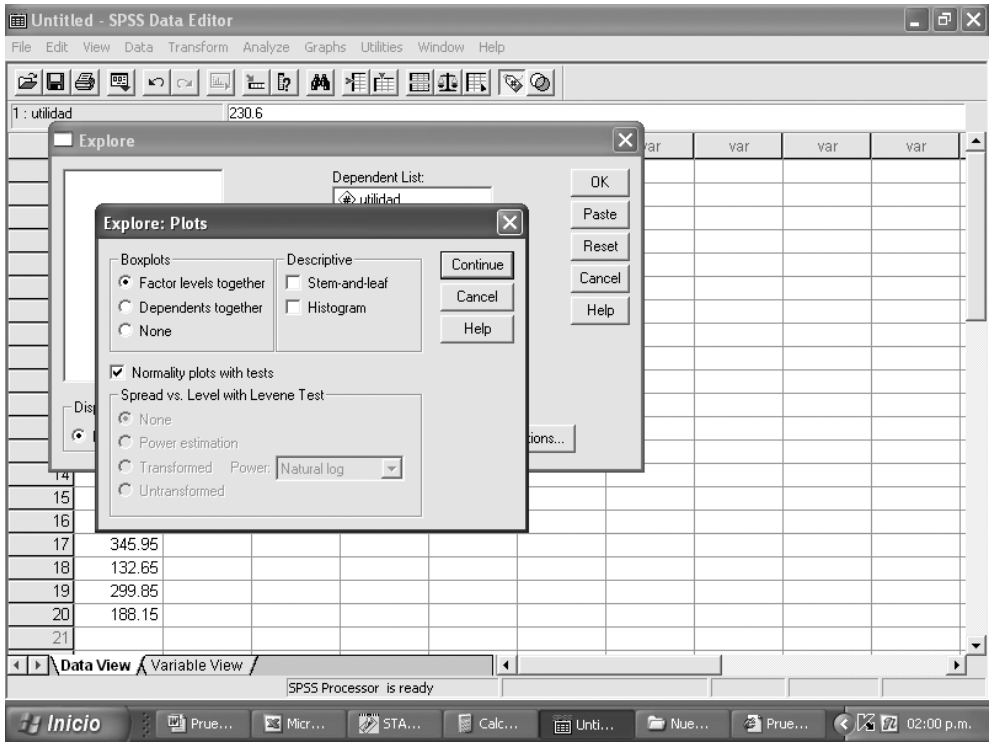
Esto está indicando, gráficamente, que efectivamente la variable sigue una distribución normal, tal y como se demostró mediante la prueba de hipótesis de Shapiro-Wilk.

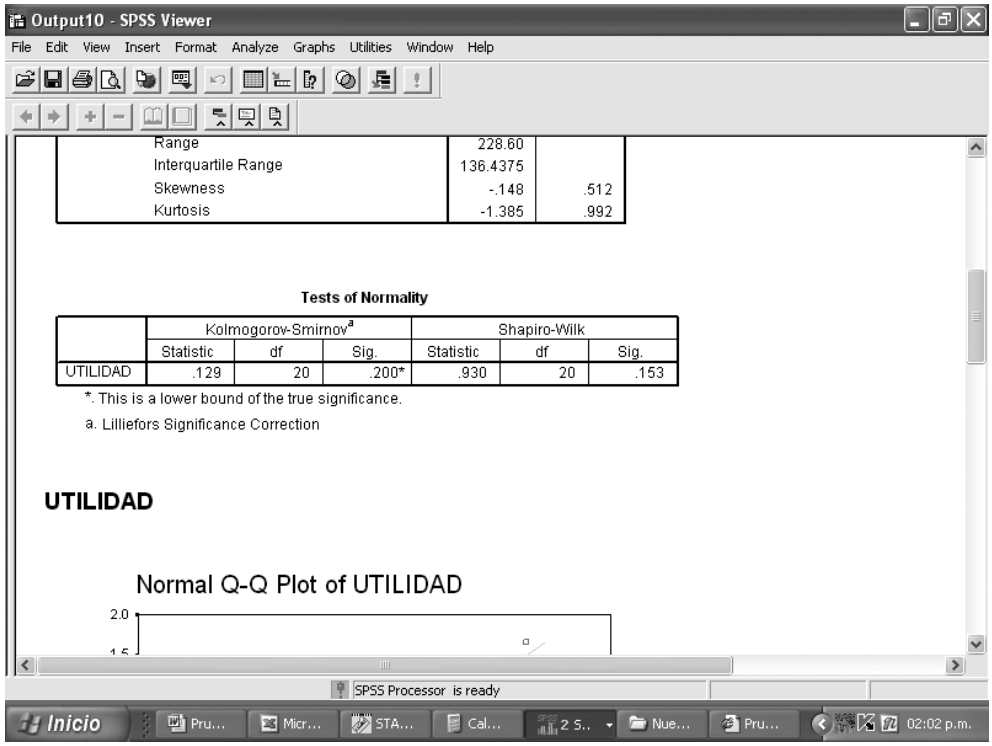
Ahora obsérvese esta misma prueba realizada con el SPSS:

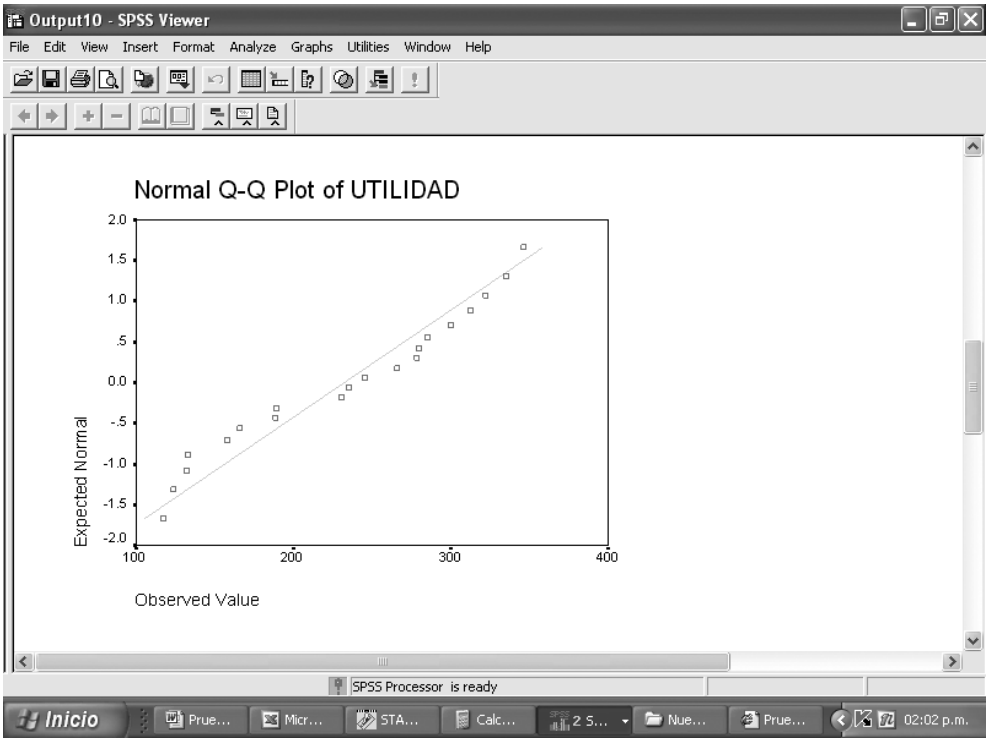



Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Frequencies...
Descriptives...
Explore...
Crosstabs...
Ratio...
1 : utilidad
utilidad
var
1 230.60
2 335.05
3 117.35
4 265.90
5 189.45
6 245.00
7 312.20
8 165.75
9 284.90
10 132.10
11 279.65
12 234.85
13 322.25
14 157.65
15 278.35
16 123.35
17 345.95
18 132.65
19 299.85
20 188.15
21
Data View
Variable View
Examine
SPSS Processor is ready
Inicio
Prue...
Micr...
STA...
Calc...
Unti...
Nue...
Prue...
01:54 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : utilidad
230.6
Explore
Dependent List:
utilidad
Factor List:
Label Cases by:
OK
Paste
Reset
Cancel
Help
Display
Both
Statistics
Plots
Statistics...
Plots...
Options...
15 278.35
16 123.35
17 345.95
18 132.65
19 299.85
20 188.15
21
Data View
Variable View
SPSS Processor is ready
Inicio
Prue...
Micr...
STA...
Calc...
Unti...
Nue...
Prue...
01:55 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : utilidad
230.6
Explore
Dependent List:
utilidad
OK
Paste
Reset
Cancel
Help
Explore: Plots
Boxplots
Factor levels together
Dependents together
None
Descriptive
Stem-and-leaf
Histogram
Continue
Cancel
Help
Normality plots with tests
Spread vs. Level with Levene Test
None
Power estimation
Transformed
Power:
Natural log
Untransformed
345.95
132.65
299.85
188.15
Data View
Variable View
SPSS Processor is ready
Inicio
02:00 p.m.


Output10 - SPSS Viewer
File Edit View Insert Format Analyze Graphs Utilities Window Help
Range 228.60
Interquartile Range 136.4375
Skewness -.148 .512
Kurtosis -1.385 .992
Tests of Normality
Kolmogorov-Smirnov[a]
Shapiro-Wilk
Statistic df Sig. Statistic df Sig.
UTILIDAD .129 20 .200* .930 20 .153
*. This is a lower bound of the true significance.
a. Lilliefors Significance Correction
UTILIDAD
Normal Q-Q Plot of UTILIDAD
2.0
SPSS Processor is ready
Inicio
02:02 p.m.



En la penúltima imagen, se observa igual resultado de la dócima de Shapiro-Wilk respecto al obtenido con el STATGRAPHICS Plus, y en la última imagen, se observa el gráfico.

5.6. Prueba de Wilcoxon.

Véase un ejemplo.

Ejemplo 4:

Un maestrante que está cursando la Maestría en Gestión Turística, se encuentra realizando la investigación que responde a su tema de tesis de grado.
En una fase de dicha investigación, ha requerido conocer, cuál ha sido el nivel de calidad percibido por los turistas respecto al servicio que ofrece el restaurante italiano del hotel. Para ello, seleccionó una muestra de clientes externos a encuestar, y a cada miembro de la misma, le administró un cuestionario conformado por 22 afirmaciones agrupadas en 5 dimensiones y empleando una escala de medición tipo Likert (5 puntos), que le permitió conocer cuál fue la percepción de los turistas acerca de la calidad del servicio

recibido. Después de procesar los datos obtenidos de la aplicación del cuestionario, pudo detectar cuáles fueron las deficiencias de la calidad del servicio percibidas por los clientes encuestados, e introdujo mejoras radicales.

Antes que los clientes externos encuestados culminaran su estancia en el hotel, el maestrante volvió a administrarles igual cuestionario, para conocer en qué medida las mejoras introducidas, lograron o no, variar la percepción de los consumidores.

Bajo un nivel de seguridad del 90%, el maestrante sospecha que cuando realice la comparación de los datos obtenidos antes de la introducción de las mejoras, respecto a los datos obtenidos después de las mismas, sí habrá diferencias considerables entre ellos, por lo cual no seguirán una distribución normal. A continuación se muestran los datos:

Dimensiones del cuestionario	Primeras observaciones	Segundas observaciones
$\mathbf{D_1}$ (elementos tangibles)	4	5
$\mathbf{D_2}$ (fiabilidad)	1	5
$\mathbf{D_3}$ (responsabilidad)	3	5
$\mathbf{D_4}$ (seguridad)	4	4
$\mathbf{D_5}$ (empatía)	2	5

Solución:

Variable de estudio X: nivel de calidad percibido del servicio (medido con una escala de 1 a 5).

H_0: $X_n - Y_n \sim N(\mu; \sigma^2)$ (la diferencia entre las primeras observaciones y las segundas, sigue una distribución normal)

H_1: $X_n - Y_n$ no$\sim N(\mu; \sigma^2)$ (la diferencia entre las primeras observaciones y las segundas, no sigue una distribución normal)

Se aclara que:

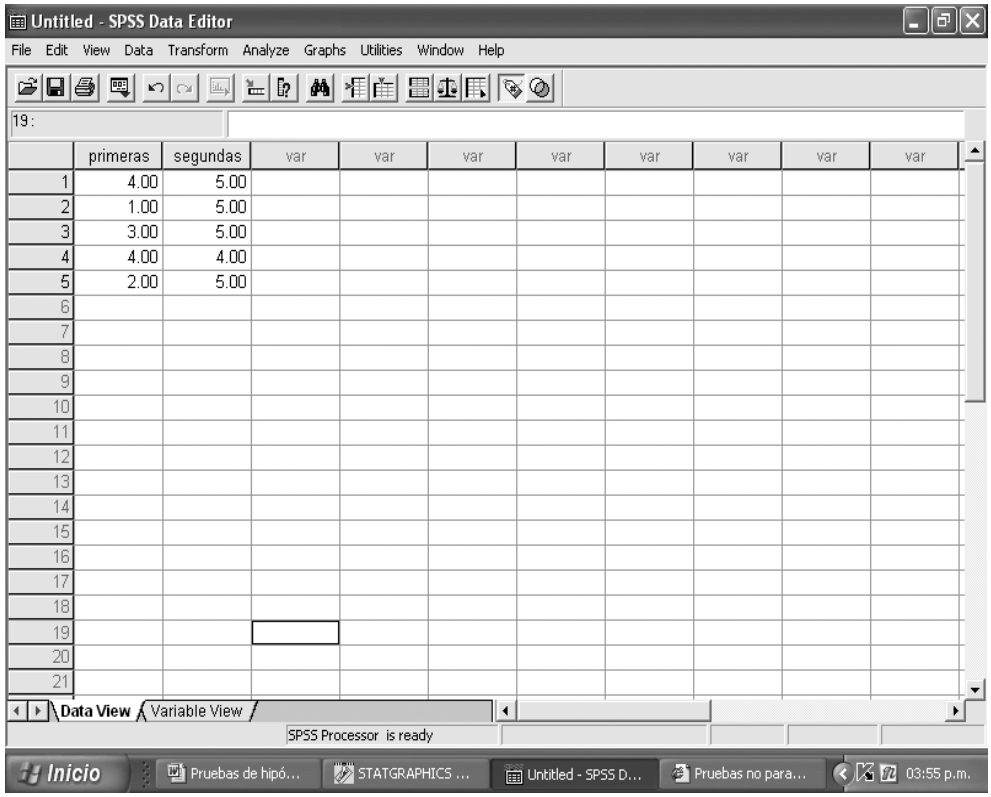
X_n: primeras observaciones (realizadas antes de introducir las mejoras de calidad).

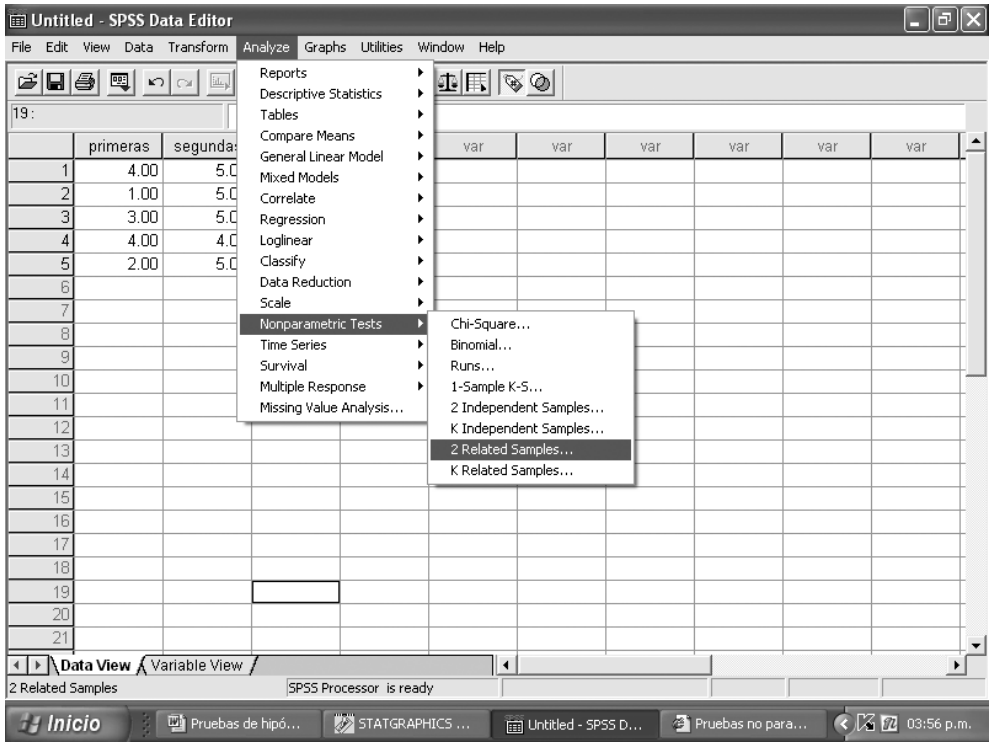
Y_n: segundas observaciones (realizadas después de introducir las mejoras de calidad).

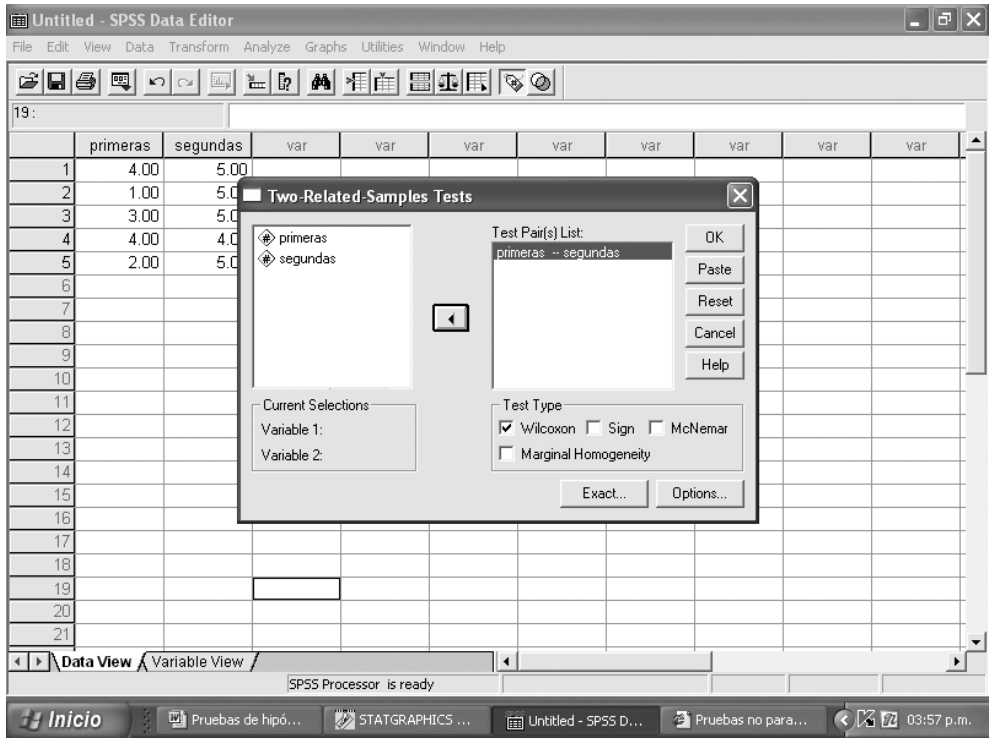
Significado de la escala empleada: **1----2----3----4----5**
TI TS

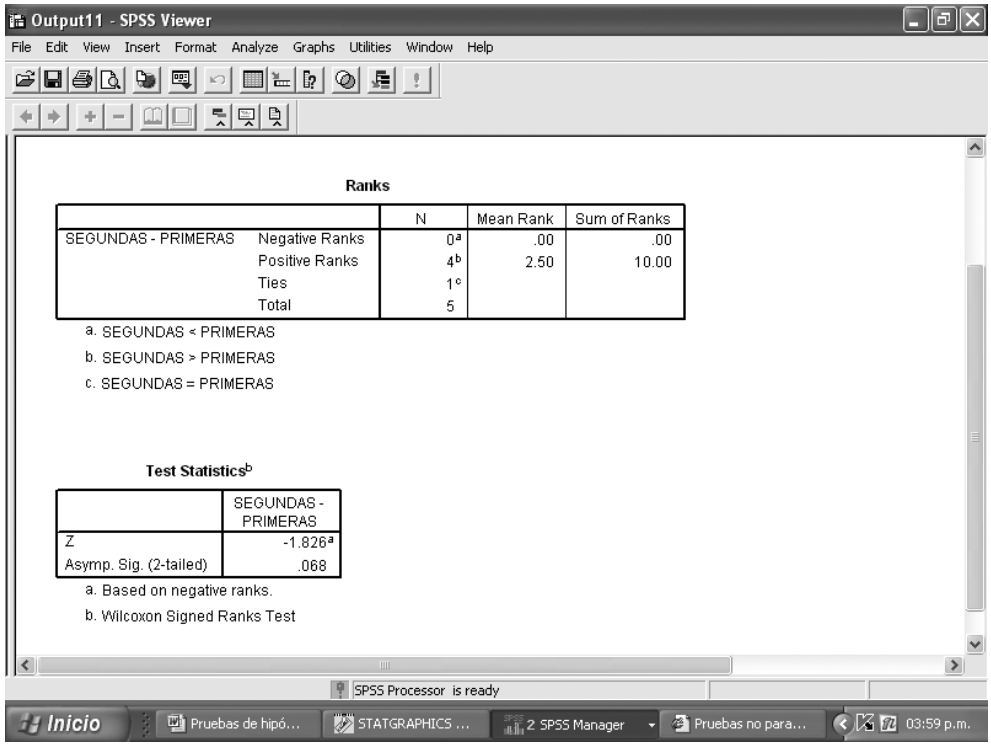
1: totalmente insatisfactorio (TI)
2: insatisfactorio
3: ni insatisfactorio ni satisfactorio
4: satisfactorio
5: totalmente satisfactorio (TS)

Utilizando el SPSS para realizar esta dócima de hipótesis no paramétrica, sería:




Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Chi-Square...
Binomial...
Runs...
1-Sample K-S...
2 Independent Samples...
K Independent Samples...
2 Related Samples...
K Related Samples...
primeras
segundas
Data View
Variable View
2 Related Samples
SPSS Processor is ready
Inicio
03:56 p.m.


Untitled - SPSS Data Editor
Two-Related-Samples Tests
primeras
segundas
Test Pair(s) List:
primeras -- segundas
OK
Paste
Reset
Cancel
Help
Current Selections
Variable 1:
Variable 2:
Test Type
Wilcoxon
Sign
McNemar
Marginal Homogeneity
Exact...
Options...
Data View
Variable View
SPSS Processor is ready
Inicio
03:57 p.m.



Según se observa, el valor de probabilidad de la dócima de Wilcoxon es igual a 0.07. Como ese valor de probabilidad es menor que 0.10, entonces sí se cumple la región crítica y se rechaza la hipótesis nula. Esto quiere decir que las diferencias entre las primeras observaciones y las segundas, son considerables o marcadas, lo cual evidencia que dichas diferencias no proceden de una distribución normal con un nivel de significación del 10%. El maestrante, entonces, ha comprobado que las mejoras introducidas, han provocado que el nivel de percepción de calidad del servicio, haya aumentado en los clientes externos.

5.7. Prueba de Mann-Whitney.

Véase un ejemplo.

Ejemplo 5:
En la Agencia de Viajes V ubicada en el polo turístico de Guardalavaca, el Departamento de Guías ha recopilado los datos de una encuesta aplicada a clientes canadienses y españoles. En los últimos 5 días fueron encuestados,

bajo un nivel de seguridad del 95%, 42 clientes canadienses e igual cantidad de clientes españoles, para conocer su nivel de satisfacción con los servicios de guiaje ofrecidos por la agencia durante las excursiones, circuitos y transfers. Los resultados se hallan a continuación:

Nacionalidad	Insatisfecho	Ni satisfecho ni insatisfecho	Satisfecho
Canadienses	12	9	21
Españoles	8	16	18

Solución:

Variable de estudio X: nivel de satisfacción de los clientes.

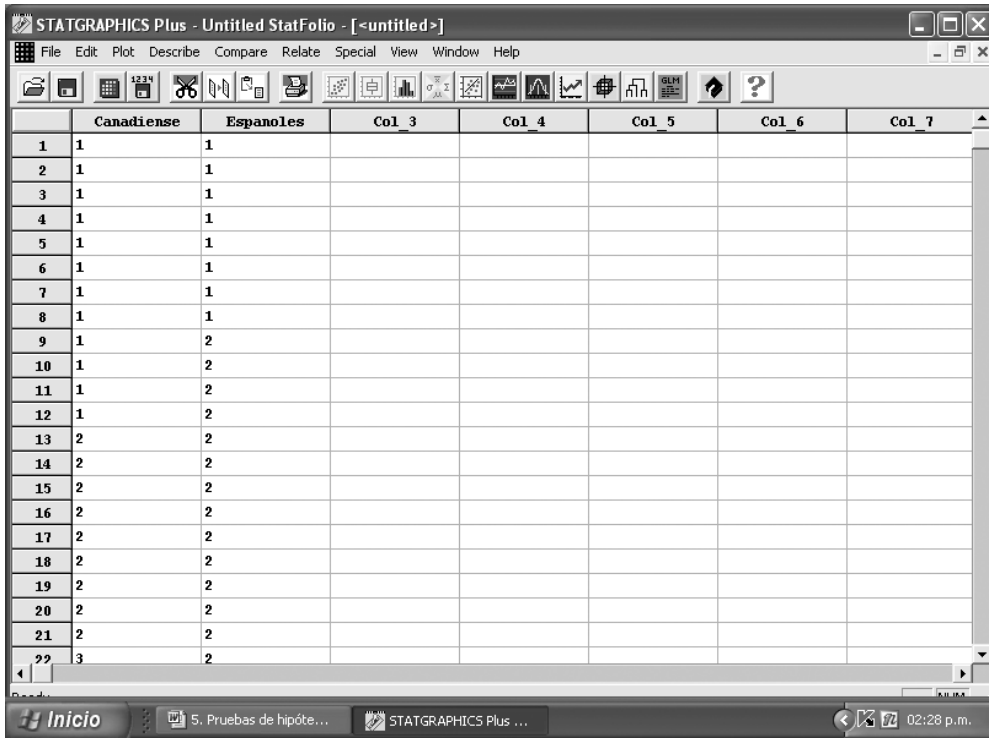
H_0: los valores de satisfacción de la muestra de clientes canadienses, son similares a los de la muestra de clientes españoles (homogeneidad)

H_1: los valores de satisfacción de una de las dos muestras de clientes, no son similares a los de la otra (heterogeneidad)

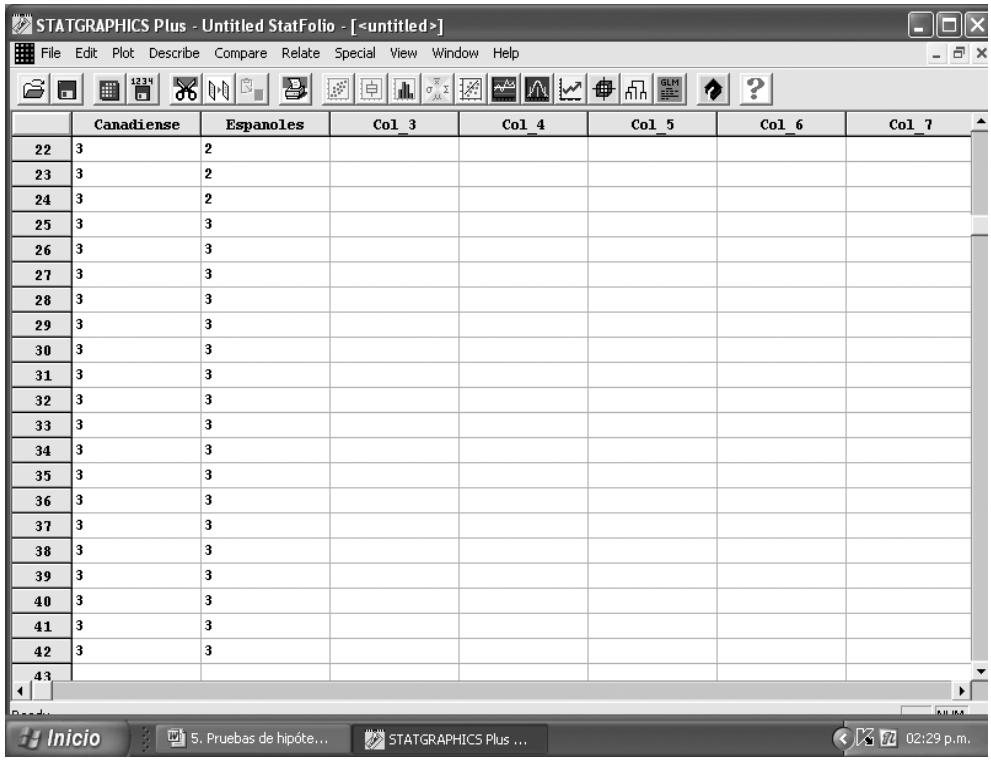
Empleando el STATGRAPHICS Plus, sería:

La variable ordinal "nivel de satisfacción" se codificará de la siguiente forma:

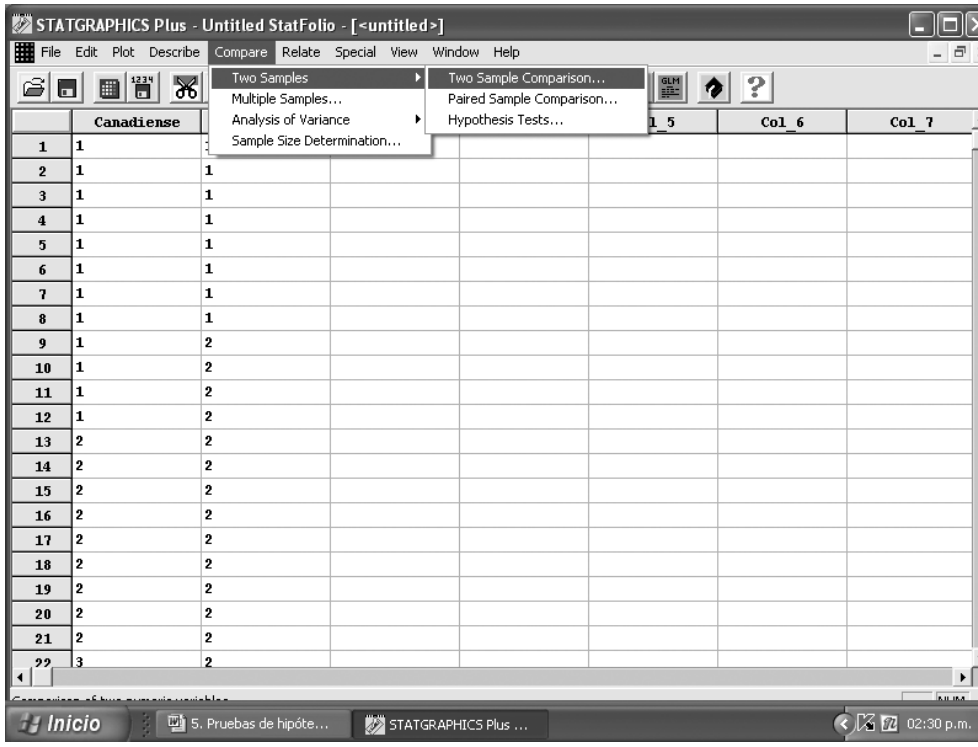
- 1: insatisfecho
- 2: ni insatisfecho ni satisfecho
- 3: satisfecho

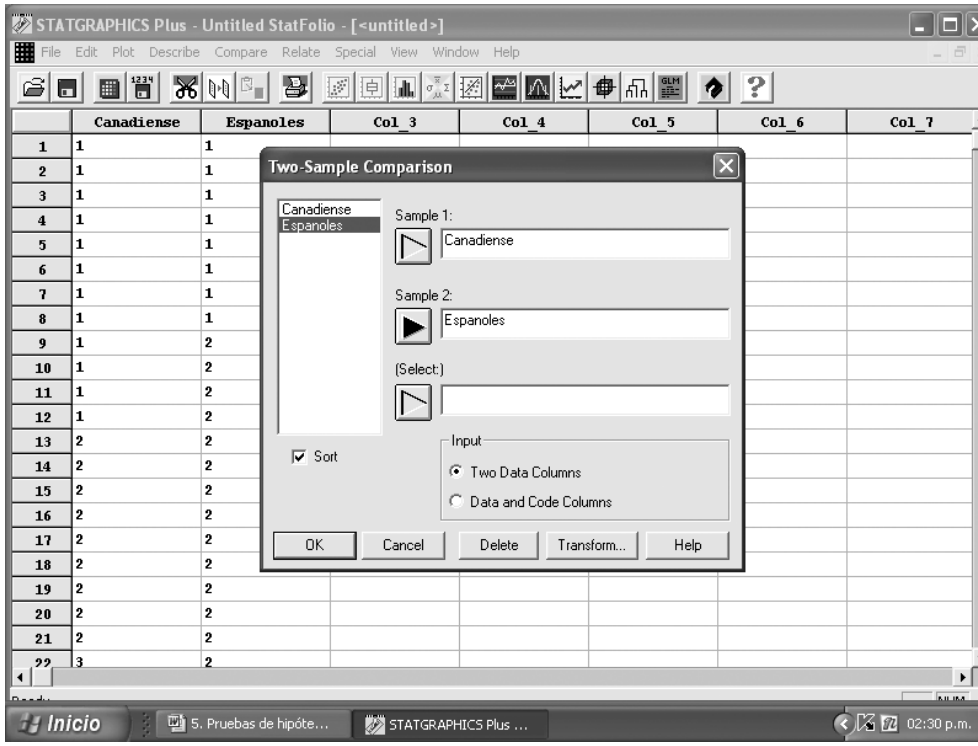


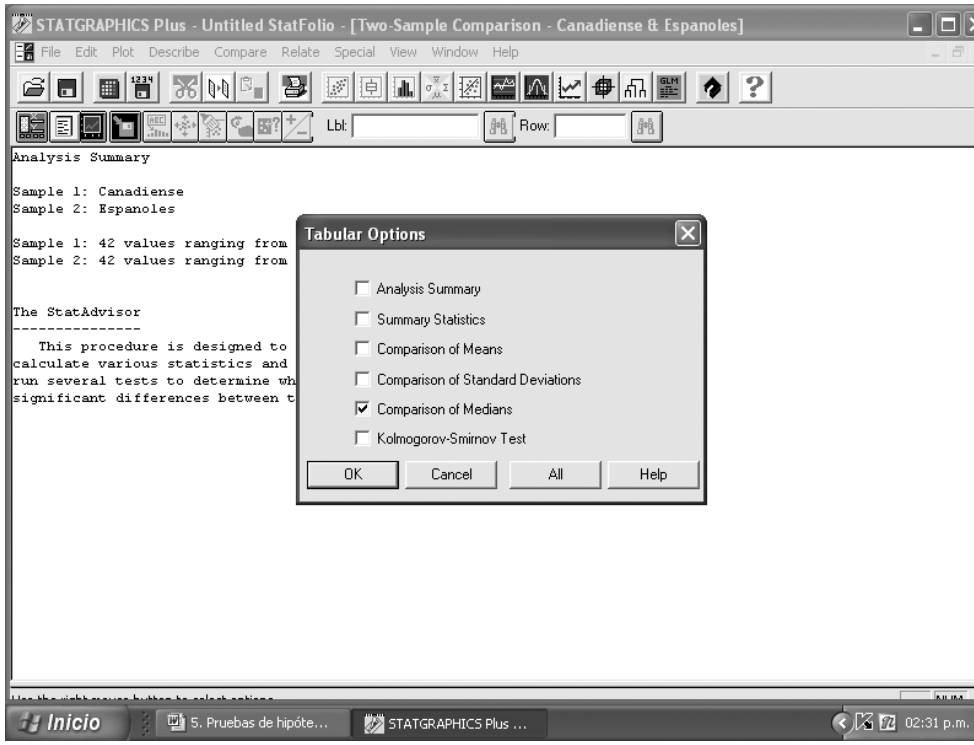
	Canadiense	Espanoles	Col_3	Col_4	Col_5	Col_6	Col_7
1	1	1					
2	1	1					
3	1	1					
4	1	1					
5	1	1					
6	1	1					
7	1	1					
8	1	1					
9	1	2					
10	1	2					
11	1	2					
12	1	2					
13	2	2					
14	2	2					
15	2	2					
16	2	2					
17	2	2					
18	2	2					
19	2	2					
20	2	2					
21	2	2					
22	3	2					

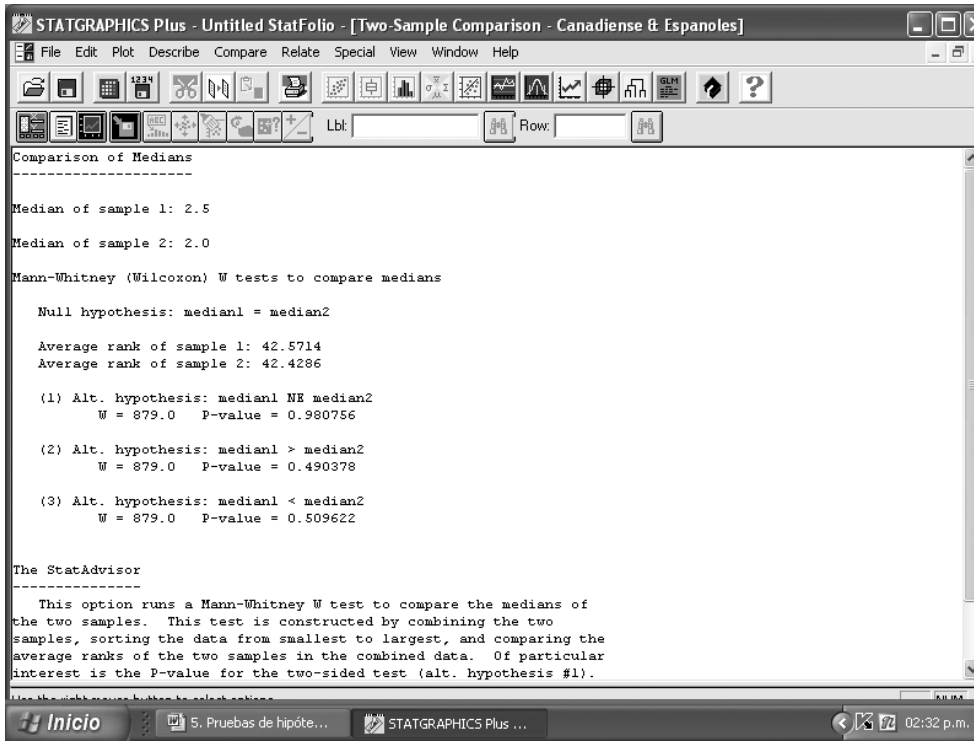


	Canadiense	Espanoles	Col_3	Col_4	Col_5	Col_6	Col_7
22	3	2					
23	3	2					
24	3	2					
25	3	3					
26	3	3					
27	3	3					
28	3	3					
29	3	3					
30	3	3					
31	3	3					
32	3	3					
33	3	3					
34	3	3					
35	3	3					
36	3	3					
37	3	3					
38	3	3					
39	3	3					
40	3	3					
41	3	3					
42	3	3					
43							


STATGRAPHICS Plus - Untitled StatFolio - [<untitled>]
File Edit Plot Describe Compare Relate Special View Window Help
Two Samples
Multiple Samples...
Analysis of Variance
Sample Size Determination...
Two Sample Comparison...
Paired Sample Comparison...
Hypothesis Tests...
Canadiense
Col_6
Col_7
Inicio
5. Pruebas de hipóte...
STATGRAPHICS Plus ...
02:30 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [<untitled>]
Canadiense
Espanoles
Col_3
Col_4
Col_5
Col_6
Col_7
Two-Sample Comparison
Canadiense
Espanoles
Sample 1:
Canadiense
Sample 2:
Espanoles
(Select:)
Sort
Input
Two Data Columns
Data and Code Columns
OK
Cancel
Delete
Transform...
Help
Inicio
5. Pruebas de hipóte...
STATGRAPHICS Plus ...
02:30 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [Two-Sample Comparison - Canadiense & Espanoles]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Analysis Summary
Sample 1: Canadiense
Sample 2: Espanoles
Sample 1: 42 values ranging from
Sample 2: 42 values ranging from
The StatAdvisor
This procedure is designed to
calculate various statistics and
run several tests to determine wh
significant differences between t
Tabular Options
Analysis Summary
Summary Statistics
Comparison of Means
Comparison of Standard Deviations
Comparison of Medians
Kolmogorov-Smirnov Test
OK
Cancel
All
Help
Inicio
5. Pruebas de hipóte...
STATGRAPHICS Plus ...
02:31 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [Two-Sample Comparison - Canadiense & Espanoles]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Comparison of Medians
Median of sample 1: 2.5
Median of sample 2: 2.0
Mann-Whitney (Wilcoxon) W tests to compare medians
Null hypothesis: median1 = median2
Average rank of sample 1: 42.5714
Average rank of sample 2: 42.4286
(1) Alt. hypothesis: median1 NE median2
W = 879.0 P-value = 0.980756
(2) Alt. hypothesis: median1 > median2
W = 879.0 P-value = 0.490378
(3) Alt. hypothesis: median1 < median2
W = 879.0 P-value = 0.509622
The StatAdvisor
This option runs a Mann-Whitney W test to compare the medians of
the two samples. This test is constructed by combining the two
samples, sorting the data from smallest to largest, and comparing the
average ranks of the two samples in the combined data. Of particular
interest is the P-value for the two-sided test (alt. hypothesis #1).
Inicio
5. Pruebas de hipóte...
STATGRAPHICS Plus ...
02:32 p.m.

Como el valor de probabilidad de la dócima es igual a 0.981 y el mismo no es menor que 0.05, entonces no se cumple la región crítica y se acepta la hipótesis nula.

El Departamento de Guías ha comprobado que después de aplicada la encuesta, los resultados reflejan que los niveles de satisfacción de los clientes canadienses, son muy similares a los de los clientes españoles, con un nivel de significación del 5%.

5.8. Prueba de Kruskal-Wallis.

Véase un ejemplo.

Ejemplo 6:

En un hotel de Pinar del Río, el jefe de Alimentos y Bebidas está realizando un ranking de los restaurantes y bares con que cuenta la instalación.

Ha seleccionado una muestra de 12 turistas al azar, y les ha pedido a cada miembro de la misma, que emita se valoración acerca de cómo percibe la calidad de los servicios ofrecidos en cada punto de consumo, sobre la base de siete criterios o atributos que se muestran a continuación:

- calidad de la oferta
- profesionalidad de los empleados
- higiene del local
- empatía de los empleados
- confort del local
- ambientación del local
- presencia física de los empleados

Los puntos de consumo que fueron analizados teniendo en cuenta los criterios anteriores, fueron:

- restaurante italiano
- bar de la playa
- restaurante chino
- restaurante francés
- bar de la piscina

- restaurante de comida internacional
- lobby bar
- restaurante de comida del mar
- bar mirador

La escala de medición empleada es la que se muestra:

1----------2----------3
Bajo Medio Alto

1: baja percepción de la calidad del servicio
2: nivel medio de percepción de la calidad del servicio
3: alta percepción de la calidad del servicio

En uno de los pasos iniciales de la realización del ranking, el jefe de Alimentos y Bebidas debe conocer si existen o no diferencias significativas entre cada uno de los puntos de consumo según los valores de percepción de los clientes acerca de la calidad del servicio, tomando en cuenta los siete criterios de evaluación.

Él sospecha que sí existen diferencias significativas con un nivel de seguridad del 95%. Obsérvese los datos tabulados:

	Evaluación del conjunto de criterios en cada punto de consumo								
	Rest. italiano	**Bar playa**	**Rest. chino**	**Rest. francés**	**Bar piscina**	**Rest. internac.**	**Lobby bar**	**Rest. del mar**	**Bar mirador**
1	3	1	2	3	3	2	3	1	2
2	1	2	2	3	2	1	1	3	1
3	2	1	2	2	3	3	2	3	1
4	2	1	2	3	2	3	2	3	2
5	3	2	3	3	2	2	2	3	2
6	1	3	1	3	2	3	1	3	3
7	3	1	3	3	1	3	1	3	2
8	3	1	3	1	1	3	2	2	3
9	3	2	2	3	3	2	3	1	2
10	2	1	2	3	2	1	3	1	2
11	1	3	2	2	1	3	3	3	2
12	3	1	3	3	1	2	3	3	1

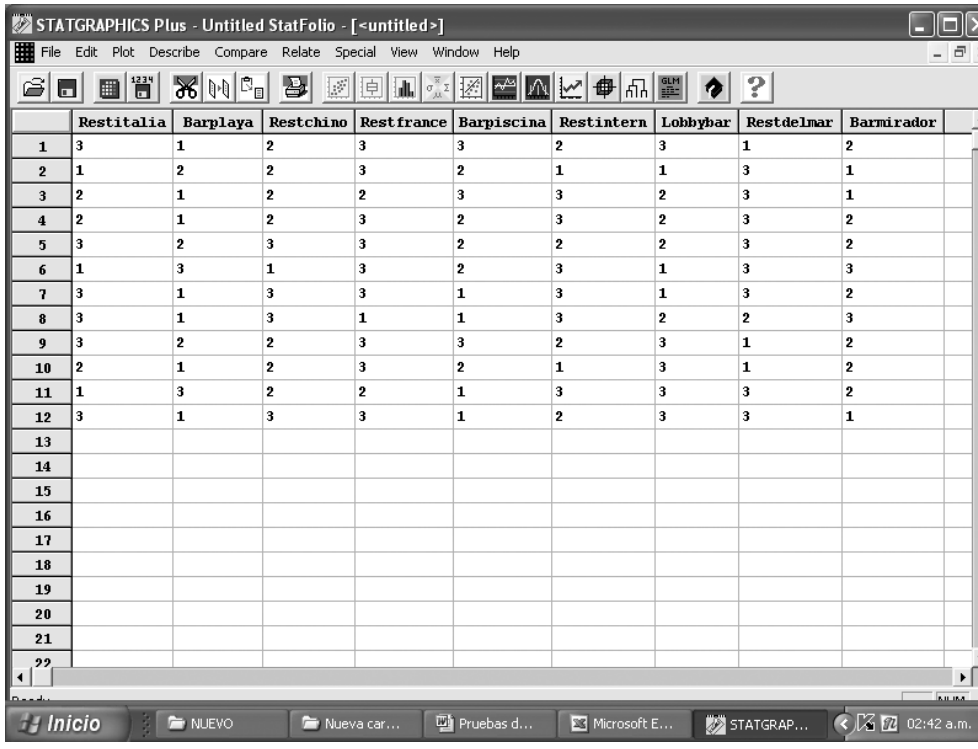
Solución:

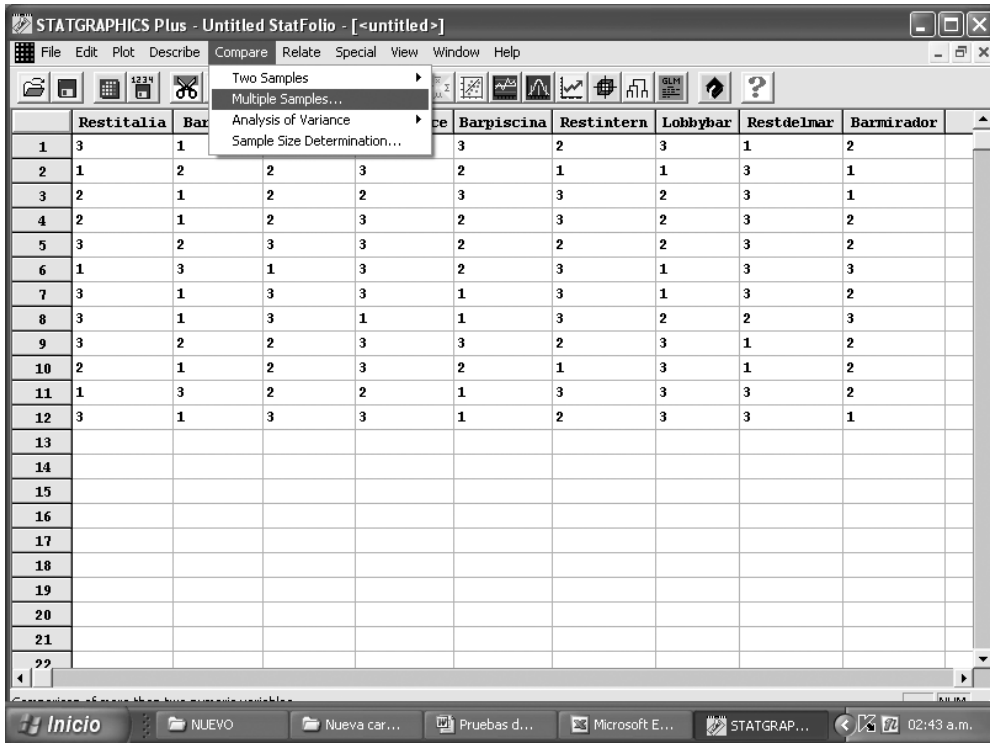
Variable de estudio X: nivel de percepción del conjunto de criterios en cada uno de los puntos de consumo.

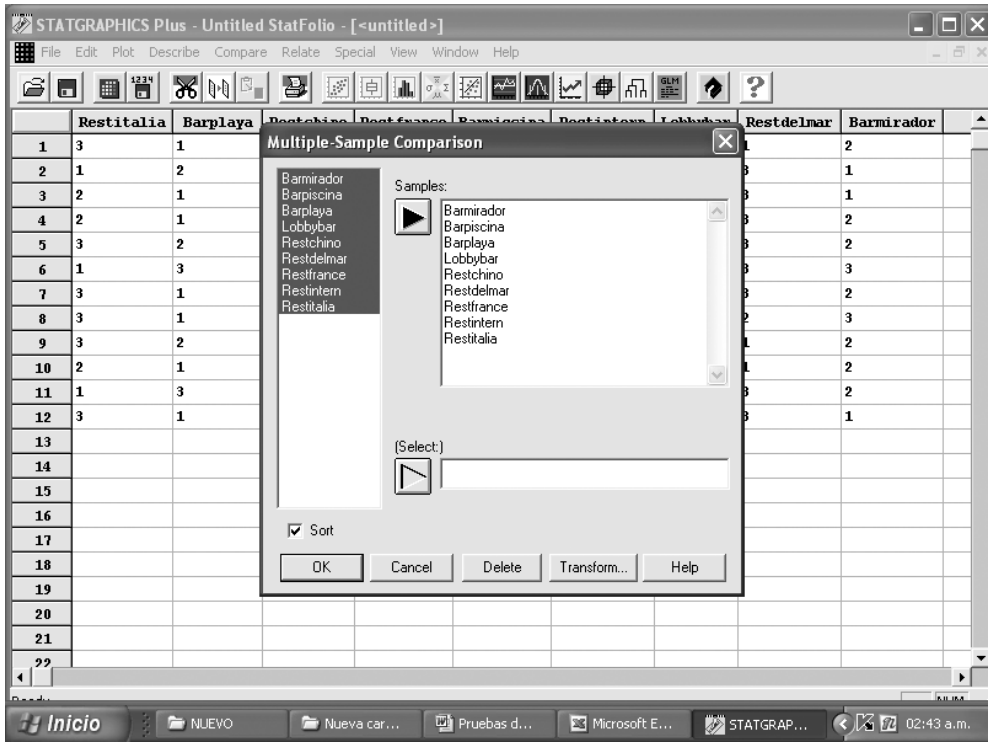
H_0: las 9 muestras provienen de la misma población ($Me_1 = Me_2 = Me_3 = Me_4 = Me_5 = Me_6 = Me_7 = Me_8 = Me_9 = Me$, o sea, no existen diferencias significativas entre las medianas de los resultados del conjunto de criterios evaluados, en cada punto de consumo)

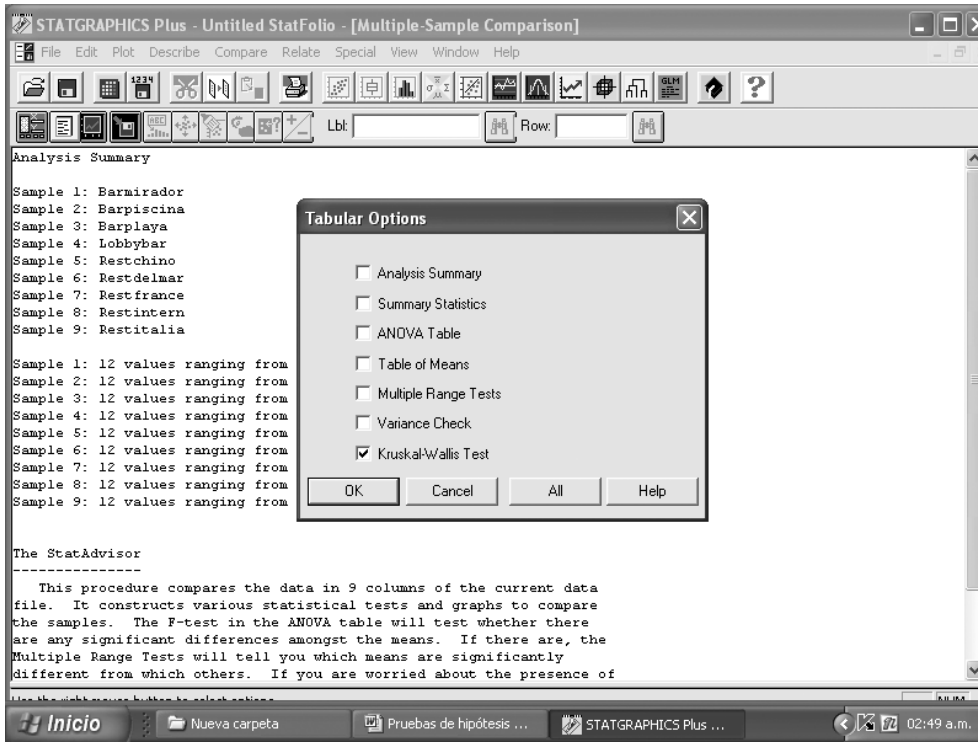
H_1: alguna muestra proviene de una población con mediana diferente a las demás (alguna $Me_i \neq Me$, o sea, sí existen diferencias significativas)

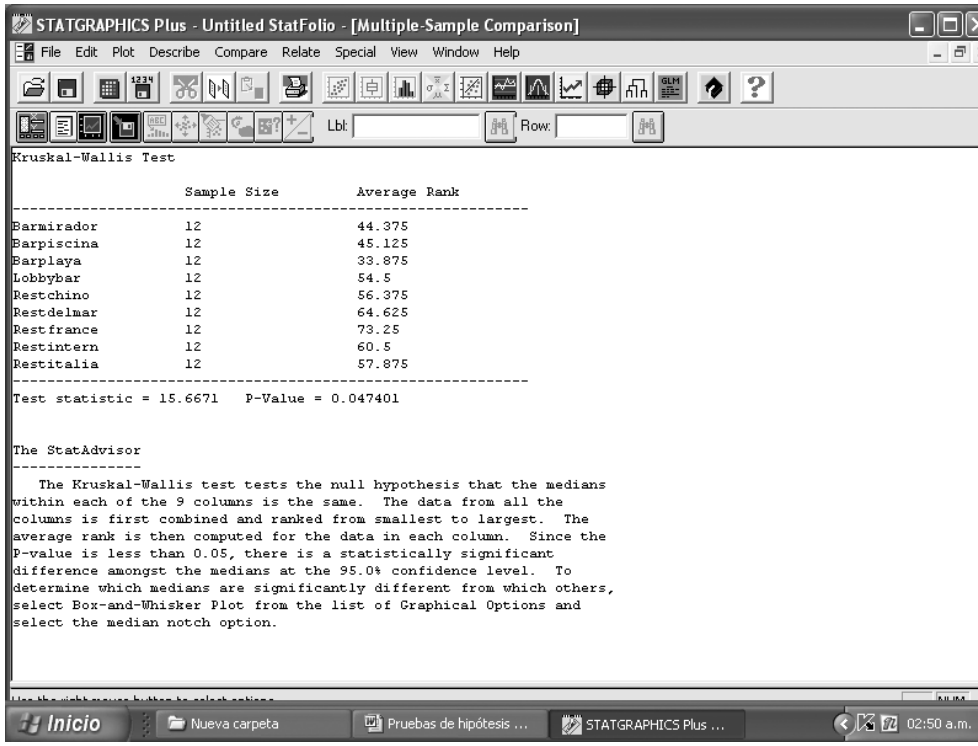
Empleando el STATGRAPHICS Plus, sería:




STATGRAPHICS Plus - Untitled StatFolio - [<untitled>]
File Edit Plot Describe Compare Relate Special View Window Help
Two Samples
Multiple Samples...
Analysis of Variance
Sample Size Determination...
Restitalia Barpiscina Restintern Lobbybar Restdelmar Barmirador
Inicio NUEVO Nueva car... Pruebas d... Microsoft E... STATGRAP... 02:43 a.m.


STATGRAPHICS Plus - Untitled StatFolio - [<untitled>]
Multiple-Sample Comparison
Barmirador
Barpiscina
Barplaya
Lobbybar
Restchino
Restdelmar
Restfrance
Restintern
Restitalia
Samples:
(Select:)
Sort
OK Cancel Delete Transform... Help
Inicio NUEVO Nueva car... Pruebas d... Microsoft E... STATGRAP... 02:43 a.m.


STATGRAPHICS Plus - Untitled StatFolio - [Multiple-Sample Comparison]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Analysis Summary
Sample 1: Barmirador
Sample 2: Barpiscina
Sample 3: Barplaya
Sample 4: Lobbybar
Sample 5: Restchino
Sample 6: Restdelmar
Sample 7: Restfrance
Sample 8: Restintern
Sample 9: Restitalia
Sample 1: 12 values ranging from
Sample 2: 12 values ranging from
Sample 3: 12 values ranging from
Sample 4: 12 values ranging from
Sample 5: 12 values ranging from
Sample 6: 12 values ranging from
Sample 7: 12 values ranging from
Sample 8: 12 values ranging from
Sample 9: 12 values ranging from
Tabular Options
Analysis Summary
Summary Statistics
ANOVA Table
Table of Means
Multiple Range Tests
Variance Check
Kruskal-Wallis Test
OK Cancel All Help
The StatAdvisor
This procedure compares the data in 9 columns of the current data file. It constructs various statistical tests and graphs to compare the samples. The F-test in the ANOVA table will test whether there are any significant differences amongst the means. If there are, the Multiple Range Tests will tell you which means are significantly different from which others. If you are worried about the presence of
Inicio Nueva carpeta Pruebas de hipótesis ... STATGRAPHICS Plus ... 02:49 a.m.

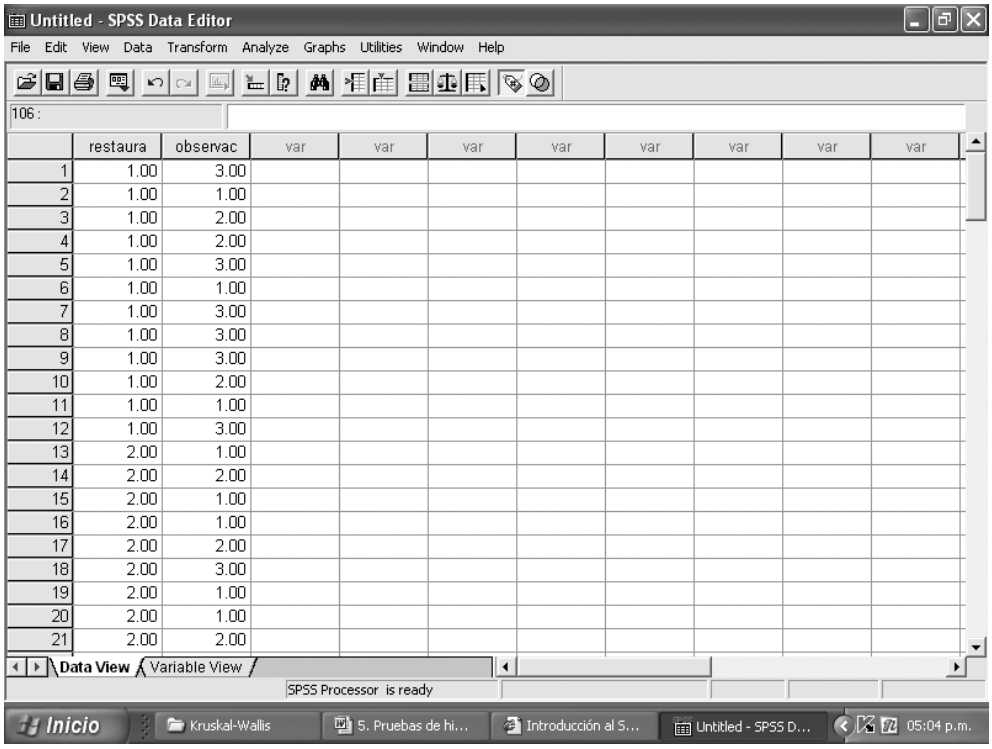

STATGRAPHICS Plus - Untitled StatFolio - [Multiple-Sample Comparison]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Kruskal-Wallis Test
Sample Size Average Rank
Barmirador 12 44.375
Barpiscina 12 45.125
Barplaya 12 33.875
Lobbybar 12 54.5
Restchino 12 56.375
Restdelmar 12 64.625
Restfrance 12 73.25
Restintern 12 60.5
Restitalia 12 57.875
Test statistic = 15.6671 P-Value = 0.047401
The StatAdvisor
The Kruskal-Wallis test tests the null hypothesis that the medians within each of the 9 columns is the same. The data from all the columns is first combined and ranked from smallest to largest. The average rank is then computed for the data in each column. Since the P-value is less than 0.05, there is a statistically significant difference amongst the medians at the 95.0% confidence level. To determine which medians are significantly different from which others, select Box-and-Whisker Plot from the list of Graphical Options and select the median notch option.
Inicio Nueva carpeta Pruebas de hipótesis ... STATGRAPHICS Plus ... 02:50 a.m.

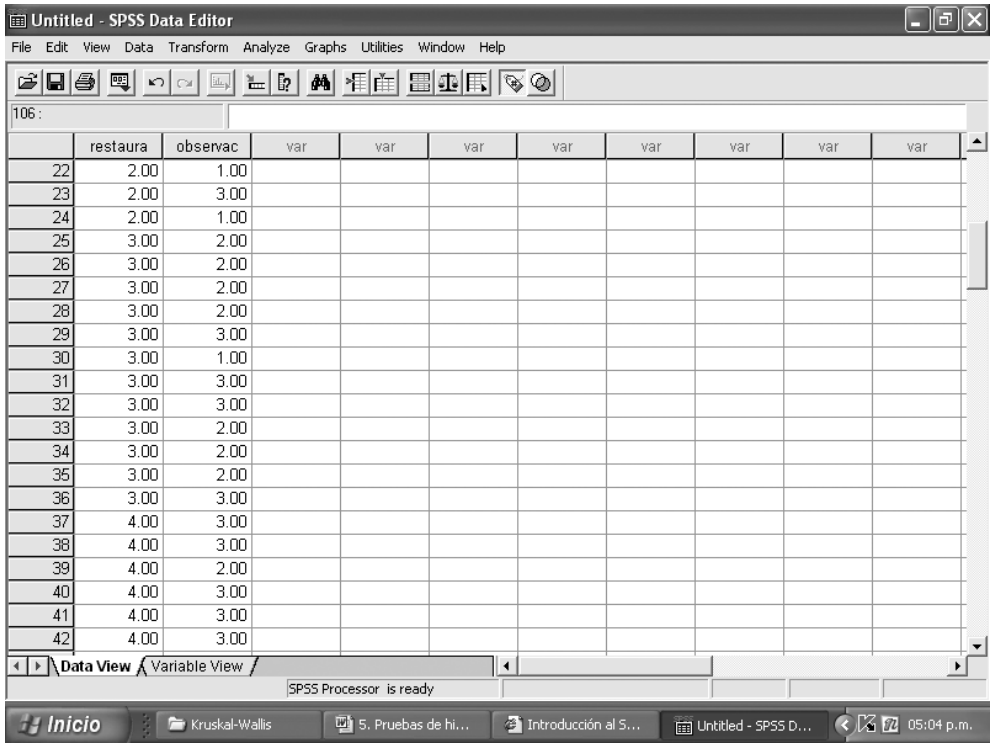
Según se observa, el valor de probabilidad de la dócima de Kruskal-Wallis es igual a 0.047. Como ese valor de probabilidad es menor que 0.05, entonces sí se cumple la región crítica y se rechaza la hipótesis nula.

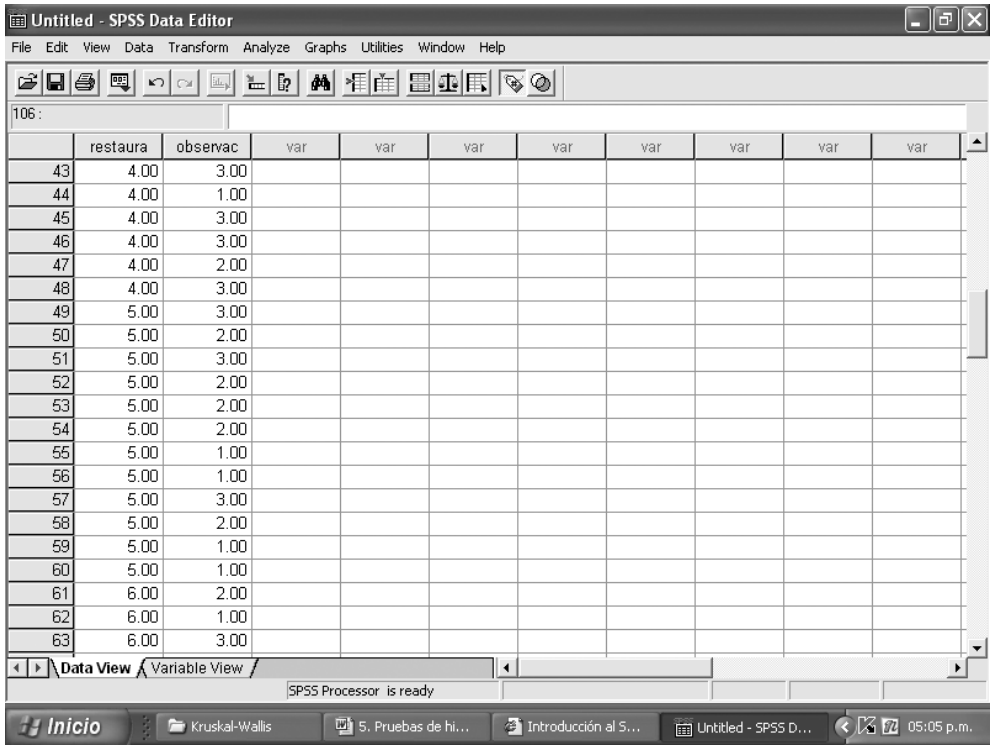
El jefe de Alimentos y Bebidas, entonces, ha corroborado que su sospecha es cierta, pues sí existen diferencias significativas entre las medianas de los resultados del conjunto de criterios evaluados, en cada punto de consumo, con un nivel de significación del 5%.

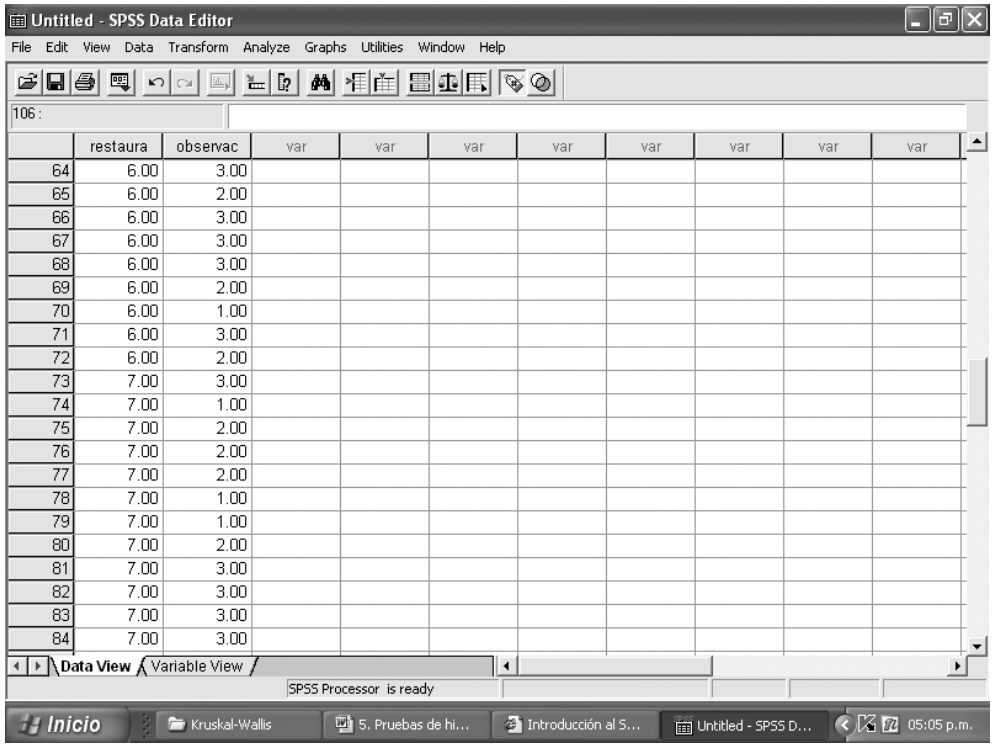
Esto significa que alguna muestra correspondiente a algún restaurante, proviene de una población con mediana diferente a las demás.

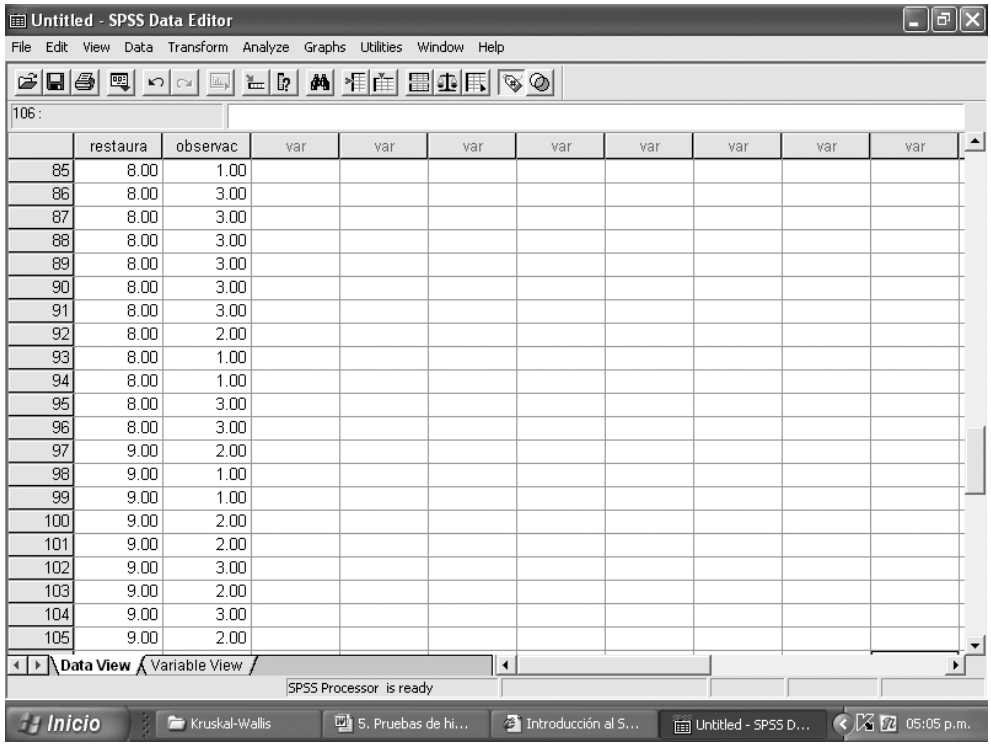
Ahora obsérvese la misma prueba mediante el SPSS:

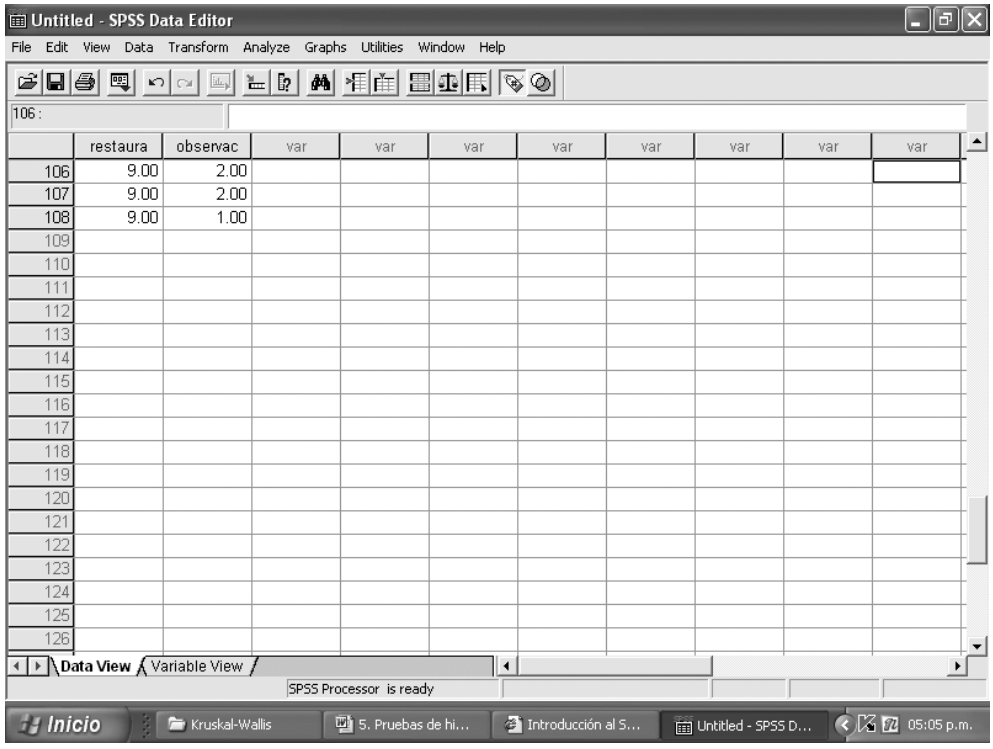


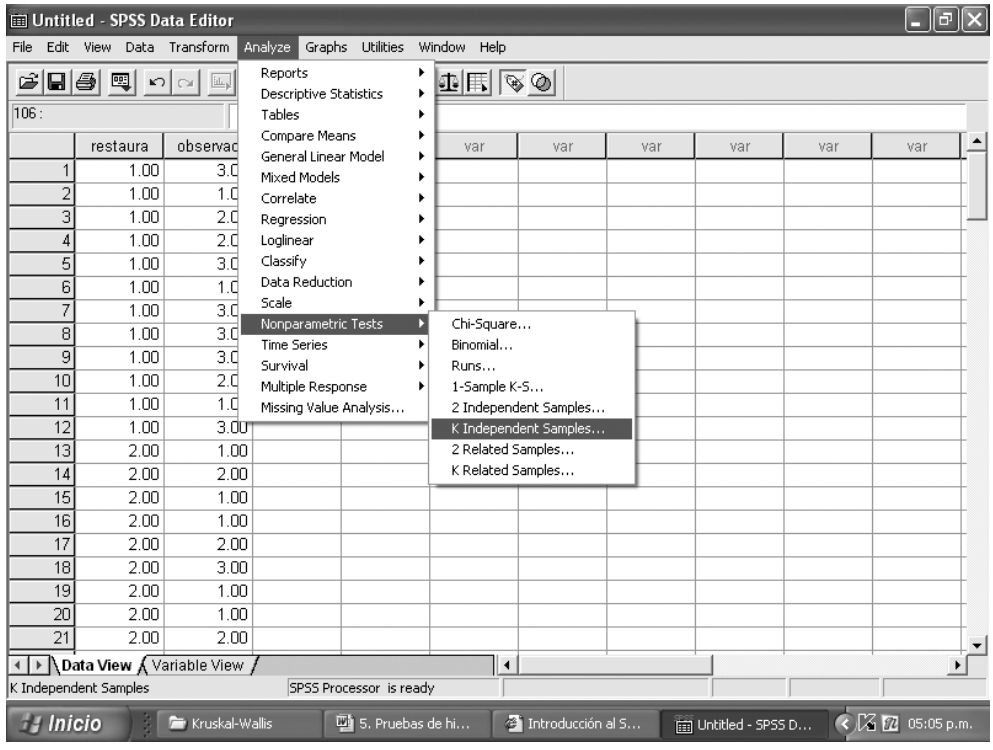

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
106 :
restaura observac var var var var var var var var
22 2.00 1.00
23 2.00 3.00
24 2.00 1.00
25 3.00 2.00
26 3.00 2.00
27 3.00 2.00
28 3.00 2.00
29 3.00 3.00
30 3.00 1.00
31 3.00 3.00
32 3.00 3.00
33 3.00 2.00
34 3.00 2.00
35 3.00 2.00
36 3.00 3.00
37 4.00 3.00
38 4.00 3.00
39 4.00 2.00
40 4.00 3.00
41 4.00 3.00
42 4.00 3.00
Data View Variable View
SPSS Processor is ready
Inicio
Kruskal-Wallis
5. Pruebas de hi...
Introducción al S...
Untitled - SPSS D...
05:04 p.m.

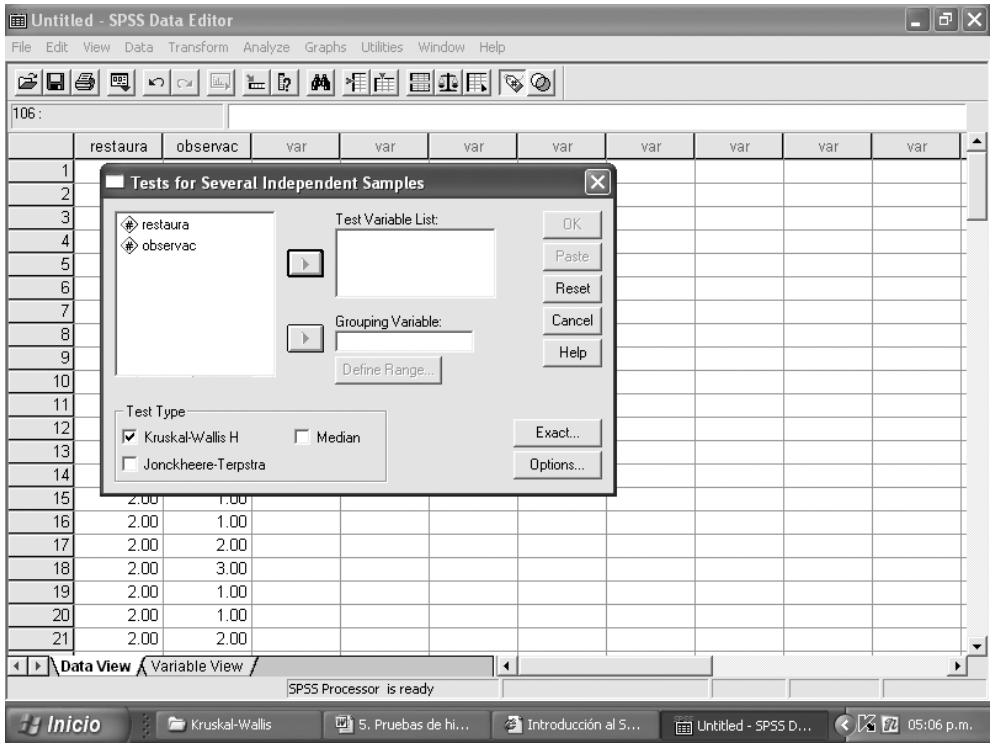

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
106 :
restaura observac var var var var var var var var
43 4.00 3.00
44 4.00 1.00
45 4.00 3.00
46 4.00 3.00
47 4.00 2.00
48 4.00 3.00
49 5.00 3.00
50 5.00 2.00
51 5.00 3.00
52 5.00 2.00
53 5.00 2.00
54 5.00 2.00
55 5.00 1.00
56 5.00 1.00
57 5.00 3.00
58 5.00 2.00
59 5.00 1.00
60 5.00 1.00
61 6.00 2.00
62 6.00 1.00
63 6.00 3.00
Data View Variable View
SPSS Processor is ready
Inicio
Kruskal-Wallis
5. Pruebas de hi...
Introducción al S...
Untitled - SPSS D...
05:05 p.m.

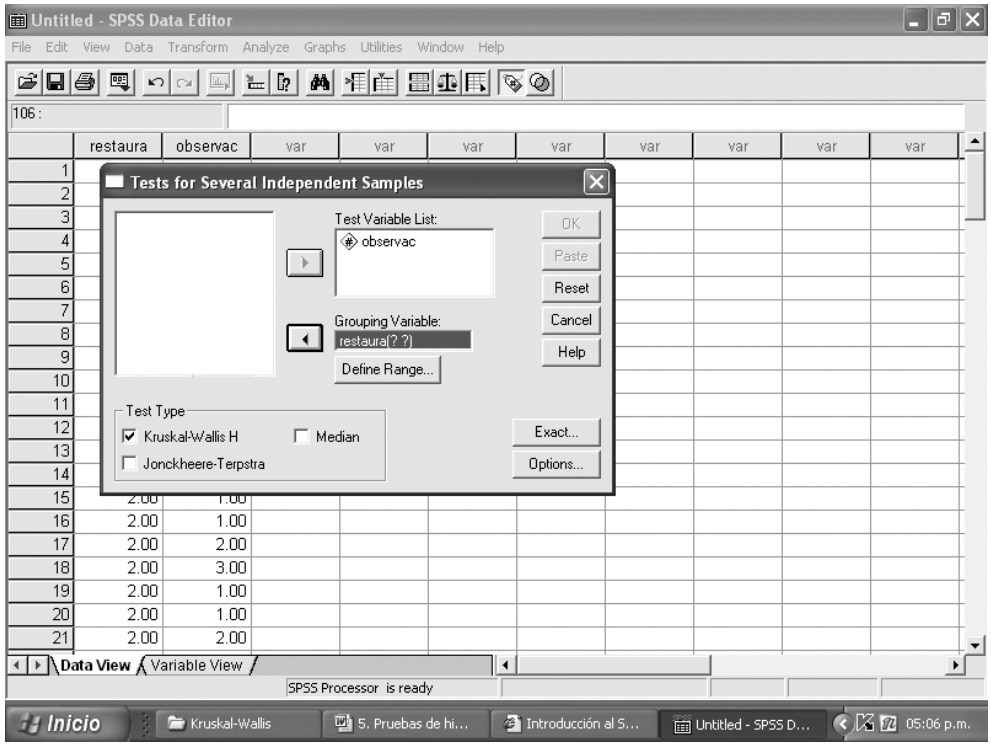

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
106 :
restaura observac var var var var var var var var
64 6.00 3.00
65 6.00 2.00
66 6.00 3.00
67 6.00 3.00
68 6.00 3.00
69 6.00 2.00
70 6.00 1.00
71 6.00 3.00
72 6.00 2.00
73 7.00 3.00
74 7.00 1.00
75 7.00 2.00
76 7.00 2.00
77 7.00 2.00
78 7.00 1.00
79 7.00 1.00
80 7.00 2.00
81 7.00 3.00
82 7.00 3.00
83 7.00 3.00
84 7.00 3.00
Data View Variable View
SPSS Processor is ready
Inicio Kruskal-Wallis 5. Pruebas de hi... Introducción al S... Untitled - SPSS D... 05:05 p.m.

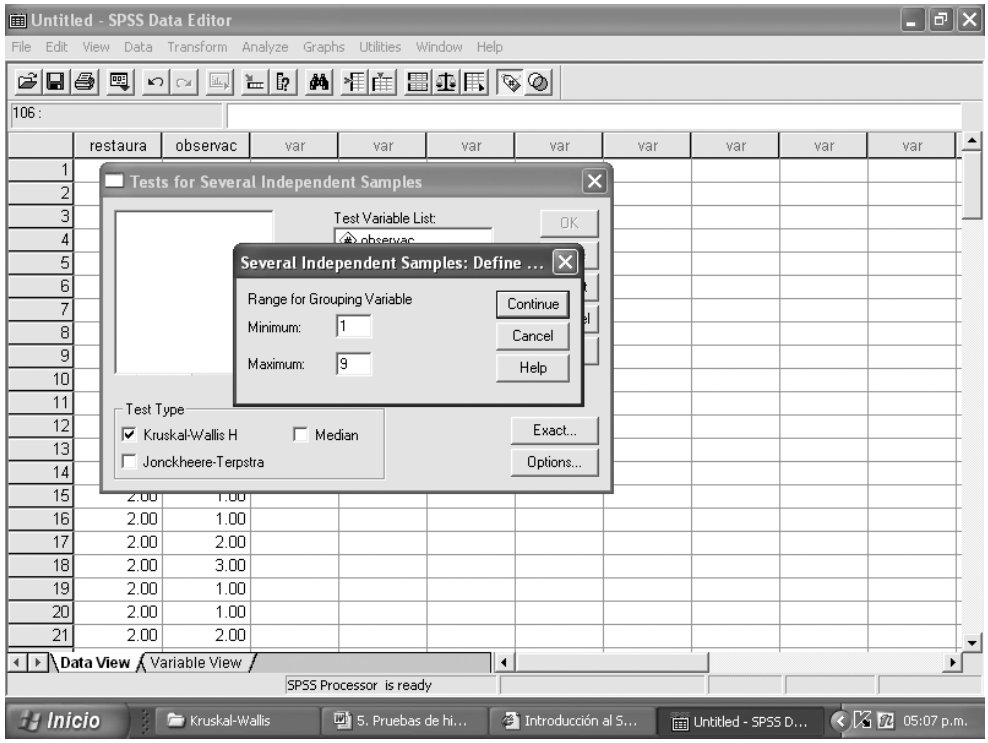

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
106 :
restaura observac var var var var var var var var
85 8.00 1.00
86 8.00 3.00
87 8.00 3.00
88 8.00 3.00
89 8.00 3.00
90 8.00 3.00
91 8.00 3.00
92 8.00 2.00
93 8.00 1.00
94 8.00 1.00
95 8.00 3.00
96 8.00 3.00
97 9.00 2.00
98 9.00 1.00
99 9.00 1.00
100 9.00 2.00
101 9.00 2.00
102 9.00 3.00
103 9.00 2.00
104 9.00 3.00
105 9.00 2.00
Data View Variable View
SPSS Processor is ready
Inicio Kruskal-Wallis 5. Pruebas de hi... Introducción al S... Untitled - SPSS D... 05:05 p.m.

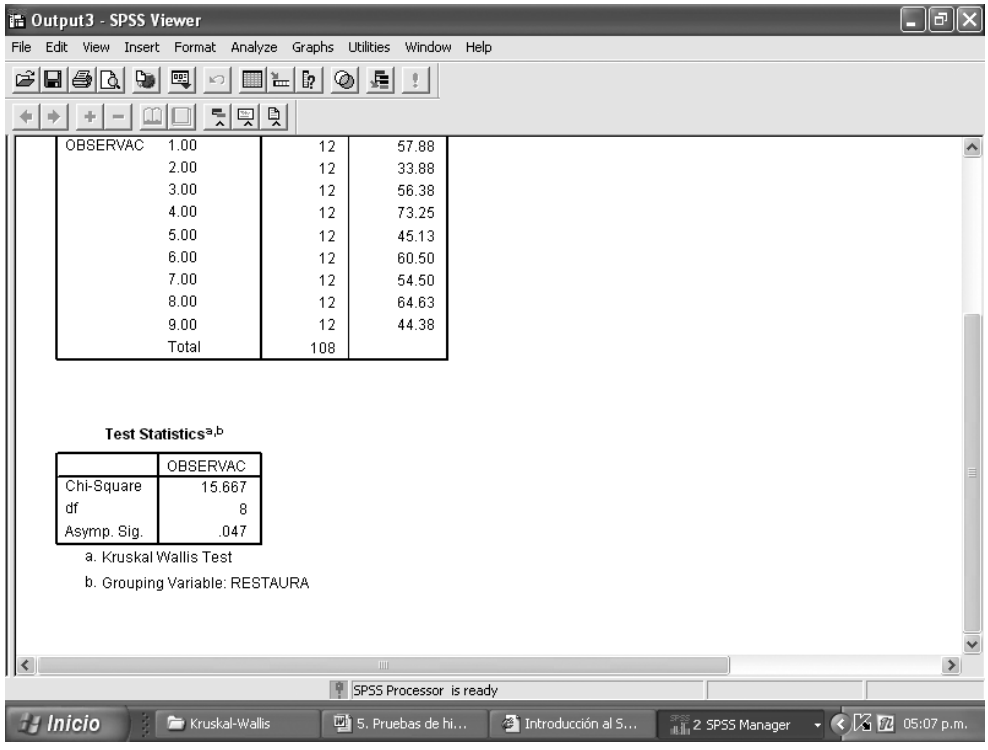

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
106 :
restaura observac var var var var var var var var
106 9.00 2.00
107 9.00 2.00
108 9.00 1.00
Data View Variable View
SPSS Processor is ready
Inicio
Kruskal-Wallis
5. Pruebas de hi...
Introducción al S...
Untitled - SPSS D...
05:05 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Chi-Square...
Binomial...
Runs...
1-Sample K-S...
2 Independent Samples...
K Independent Samples...
2 Related Samples...
K Related Samples...
106 :
restaura observac
Data View Variable View
K Independent Samples
SPSS Processor is ready
Inicio
Kruskal-Wallis
5. Pruebas de hi...
Introducción al S...
Untitled - SPSS D...
05:05 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
106 :
restaura
observac
var
Tests for Several Independent Samples
restaura
observac
Test Variable List:
Grouping Variable:
Define Range...
OK
Paste
Reset
Cancel
Help
Test Type
Kruskal-Wallis H
Median
Jonckheere-Terpstra
Exact...
Options...
2.00 1.00
2.00 2.00
2.00 3.00
2.00 1.00
2.00 1.00
2.00 2.00
Data View
Variable View
SPSS Processor is ready
Inicio
Kruskal-Wallis
5. Pruebas de hi...
Introducción al S...
Untitled - SPSS D...
05:06 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
106 :
restaura
observac
var
Tests for Several Independent Samples
Test Variable List:
observac
Grouping Variable:
restaura(? ?)
Define Range...
OK
Paste
Reset
Cancel
Help
Test Type
Kruskal-Wallis H
Median
Jonckheere-Terpstra
Exact...
Options...
2.00 1.00
2.00 2.00
2.00 3.00
2.00 1.00
2.00 1.00
2.00 2.00
Data View
Variable View
SPSS Processor is ready
Inicio
Kruskal-Wallis
5. Pruebas de hi...
Introducción al S...
Untitled - SPSS D...
05:06 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
106 :
restaura observac var var var var var var var var
Tests for Several Independent Samples
Test Variable List:
OK
observac
Several Independent Samples: Define ...
Range for Grouping Variable
Minimum: 1
Maximum: 9
Continue
Cancel
Help
Test Type
Kruskal-Wallis H
Median
Jonckheere-Terpstra
Exact...
Options...
2.00 1.00
2.00 2.00
2.00 3.00
2.00 1.00
2.00 1.00
2.00 2.00
Data View
Variable View
SPSS Processor is ready
Inicio
Kruskal-Wallis
5. Pruebas de hi...
Introducción al S...
Untitled - SPSS D...
05:07 p.m.


Output3 - SPSS Viewer
File Edit View Insert Format Analyze Graphs Utilities Window Help
OBSERVAC 1.00 12 57.88
2.00 12 33.88
3.00 12 56.38
4.00 12 73.25
5.00 12 45.13
6.00 12 60.50
7.00 12 54.50
8.00 12 64.63
9.00 12 44.38
Total 108
Test Statistics[a,b]
OBSERVAC
Chi-Square 15.667
df 8
Asymp. Sig. .047
a. Kruskal Wallis Test
b. Grouping Variable: RESTAURA
SPSS Processor is ready
Inicio
Kruskal-Wallis
5. Pruebas de hi...
Introducción al S...
2 SPSS Manager
05:07 p.m.

5.9. Prueba de independencia X^2 empleando tablas de contingencia.

Véase un ejemplo.

Ejemplo 7:

En el Hotel C ubicado en el polo turístico de Varadero, los trabajadores han pedido en Asambleas de Representantes, que se les informe mensualmente acerca de los principales resultados económicos y comerciales del hotel, para mantenerse informados.

El director general ha indicado a cada jefe de área, que dé a conocer mensualmente a sus empleados en las asambleas sindicales, los resultados básicos de las operaciones de la entidad.
Después de seis meses, ha decidido conocer si los trabajadores se encuentran o no informados, y como el hotel cuenta con 460 trabajadores, el director indagó en sólo tres áreas, pero ciertamente las que más empleados poseen en la plantilla:

Alimentos y Bebidas (150 trabajadores), Regiduría de Pisos (90) y Mantenimiento (100). Los resultados tabulados se observan a continuación:

Área	Sí	No
Alimentos y Bebidas	92	58
Regiduría de Pisos	63	27
Mantenimiento	83	17

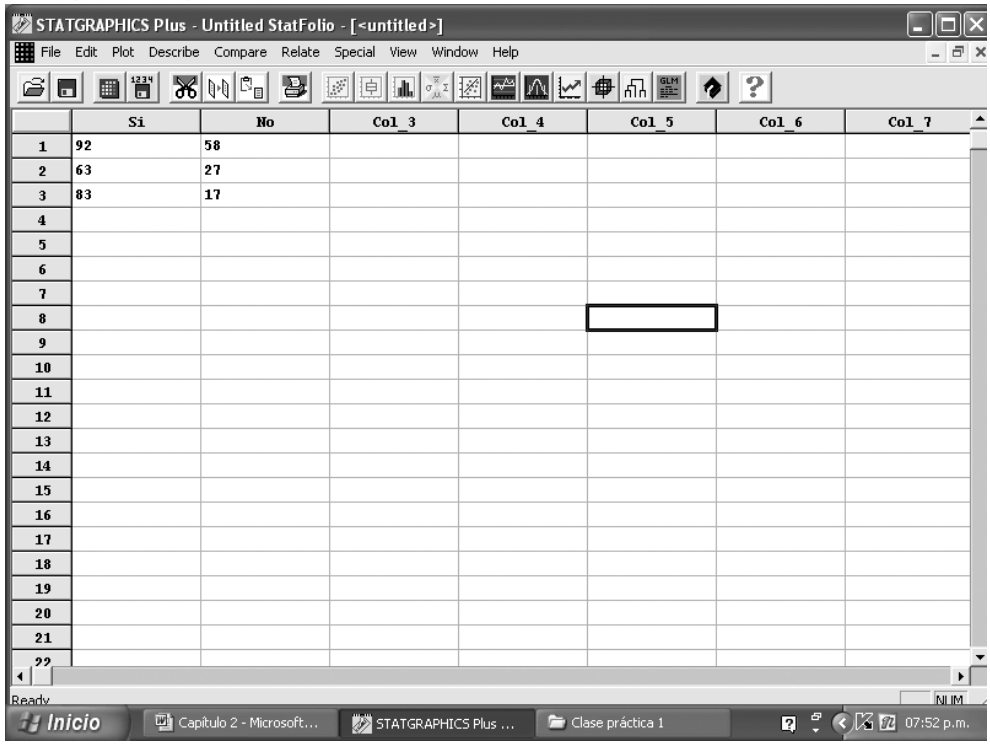
El director desea comprobar con un nivel de significación del 1% si, el hecho de que los empleados se encuentren o no informados, depende del área a que pertenecen.

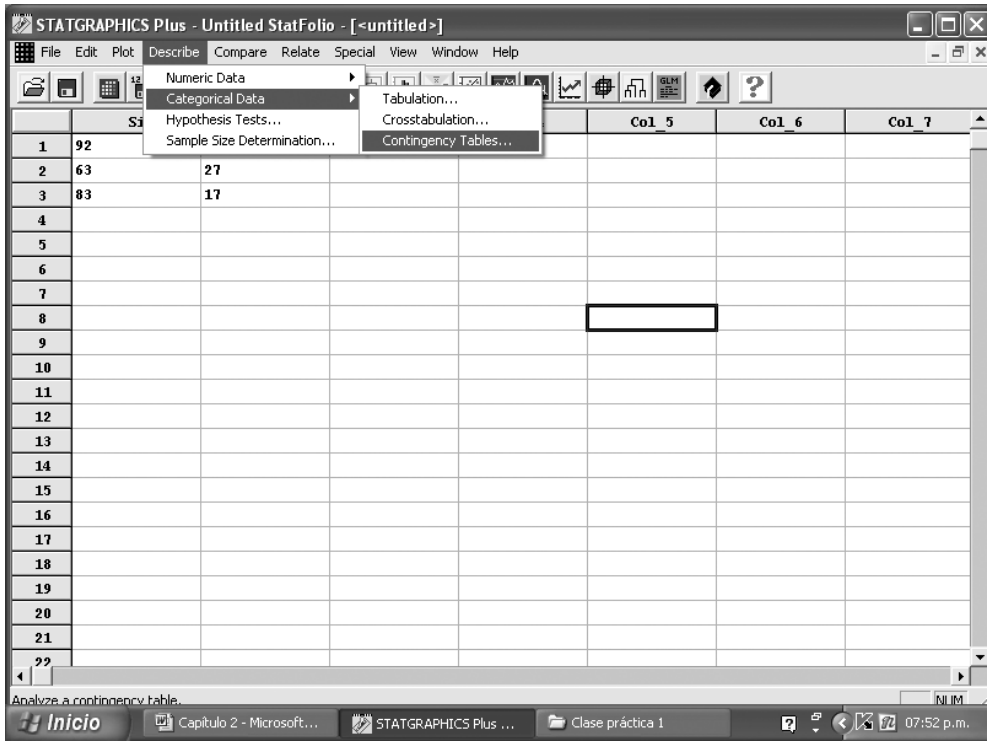
Solución:

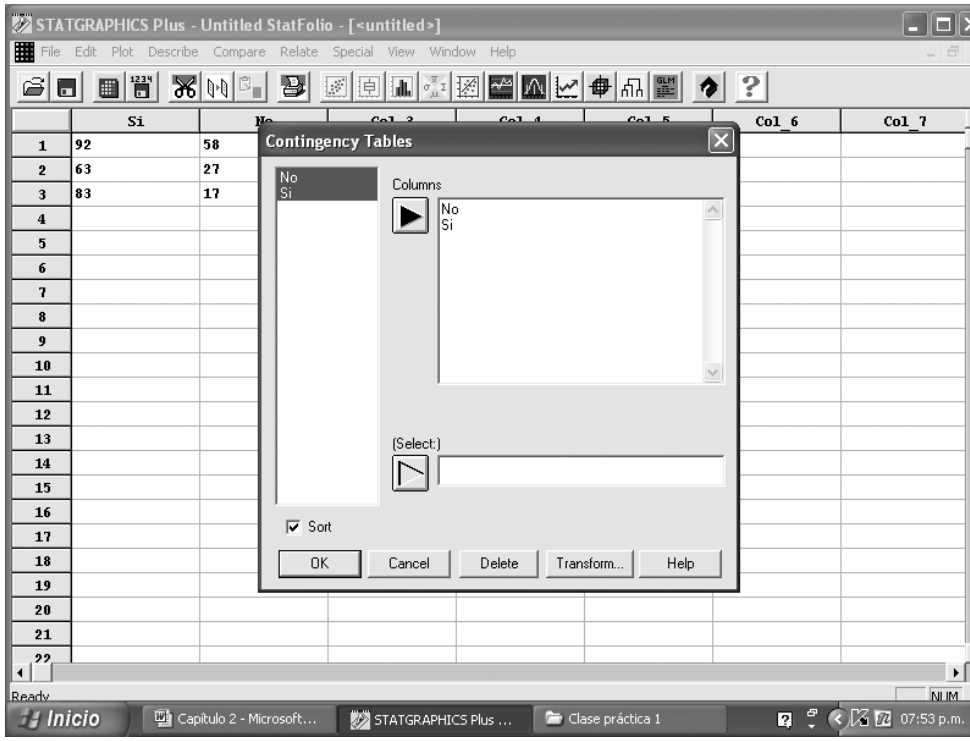
Variable de estudio X: nivel de información acerca de los principales resultados económicos y comerciales del hotel.

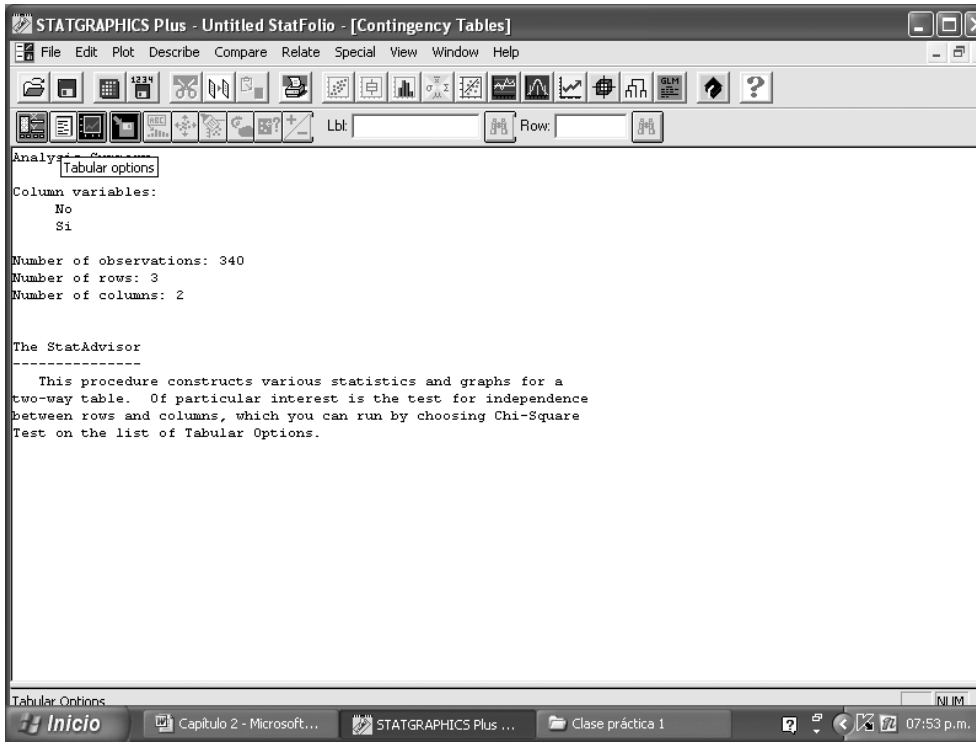
H_0: $P_{ij} = P_i * P_j$ (existe independencia entre el nivel de información y el área)

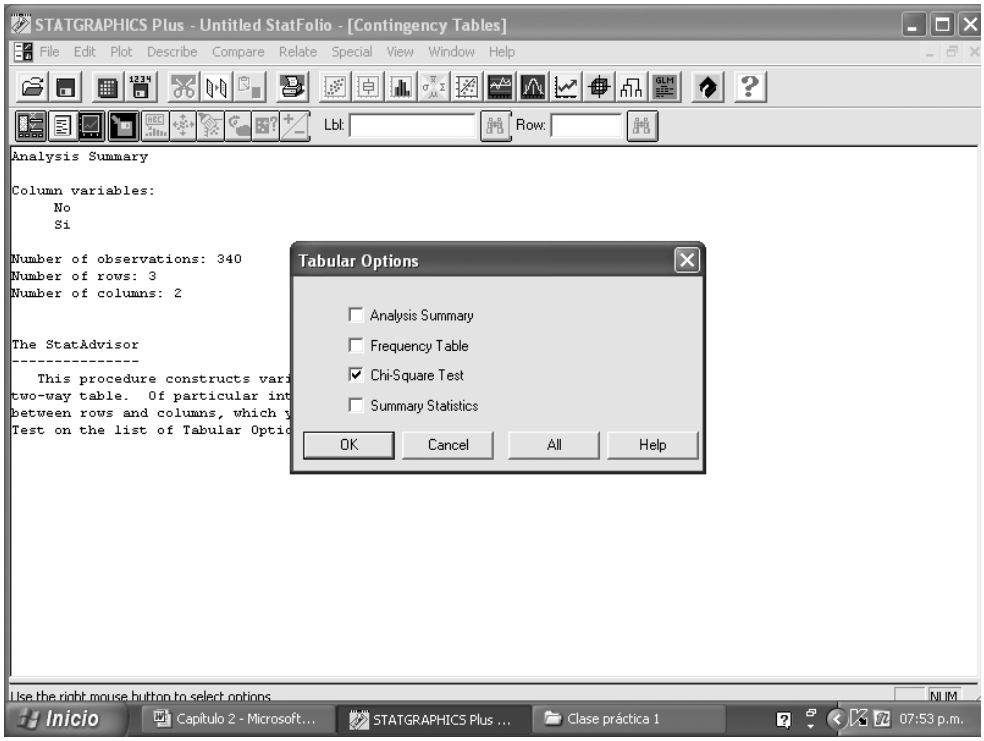
H_1: $P_{ij} \neq P_i * P_j$ (existe dependencia entre el nivel de información y el área)

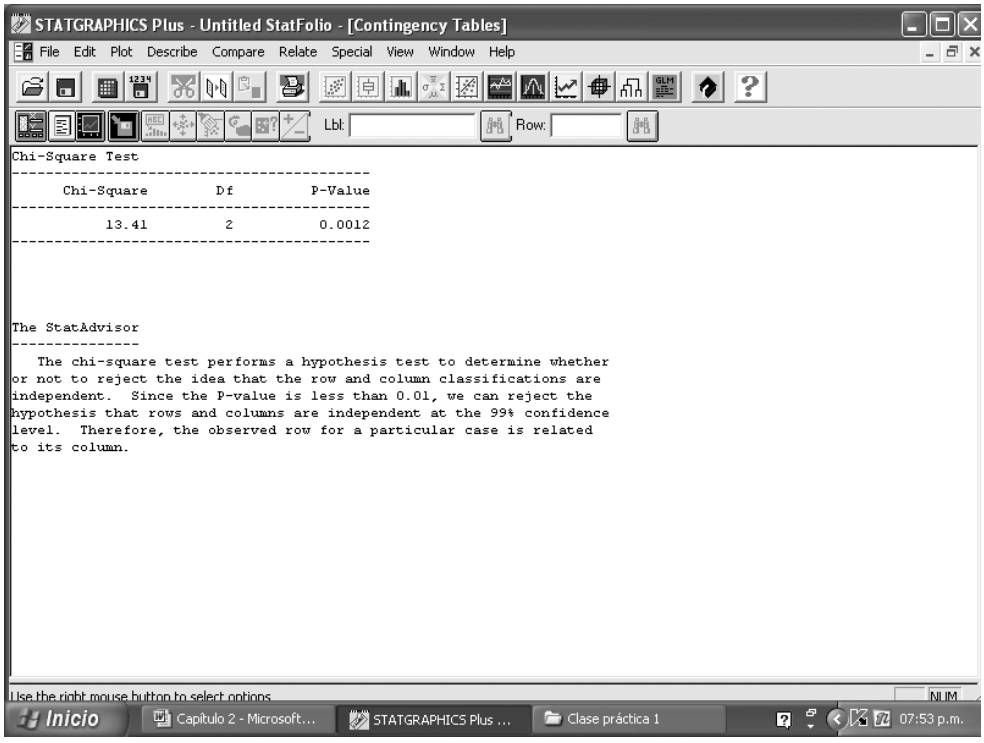





STATGRAPHICS Plus - Untitled StatFolio - [<untitled>]
File Edit Plot Describe Compare Relate Special View Window Help
Si
Col_6
Col_7
1 92 58
2 63 27
3 83 17
Contingency Tables
No
Si
Columns
No
Si
(Select:)
Sort
OK
Cancel
Delete
Transform...
Help
Ready
NUM
Inicio
Capítulo 2 - Microsoft...
STATGRAPHICS Plus ...
Clase práctica 1
07:53 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [Contingency Tables]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl:
Row:
Tabular options
Column variables:
No
Si
Number of observations: 340
Number of rows: 3
Number of columns: 2
The StatAdvisor
This procedure constructs various statistics and graphs for a
two-way table. Of particular interest is the test for independence
between rows and columns, which you can run by choosing Chi-Square
Test on the list of Tabular Options.
Tabular Options
NUM
Inicio
Capítulo 2 - Microsoft...
STATGRAPHICS Plus ...
Clase práctica 1
07:53 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [Contingency Tables]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Analysis Summary
Column variables:
No
Si
Number of observations: 340
Number of rows: 3
Number of columns: 2
The StatAdvisor
Tabular Options
Analysis Summary
Frequency Table
Chi-Square Test
Summary Statistics
OK Cancel All Help
Use the right mouse button to select options
NUM
Inicio
Capítulo 2 - Microsoft...
STATGRAPHICS Plus ...
Clase práctica 1
07:53 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [Contingency Tables]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Chi-Square Test
Chi-Square Df P-Value
13.41 2 0.0012
The StatAdvisor
The chi-square test performs a hypothesis test to determine whether or not to reject the idea that the row and column classifications are independent. Since the P-value is less than 0.01, we can reject the hypothesis that rows and columns are independent at the 99% confidence level. Therefore, the observed row for a particular case is related to its column.
Use the right mouse button to select options
NUM
Inicio
Capítulo 2 - Microsoft...
STATGRAPHICS Plus ...
Clase práctica 1
07:53 p.m.

Como el valor de probabilidad de la dócima es igual a 0.001 y el mismo es menor que 0.01, entonces sí se cumple la región crítica y se rechaza la hipótesis nula.

El director del hotel ha comprobado que el hecho de estar o no informados los trabajadores acerca de los resultados económicos y comerciales, sí depende del área a que ellos pertenecen, con un nivel del confiabilidad del 99%.

Nota: cabría pensar que las áreas donde la minoría de los empleados es la que se halla informada, están realizando un mal trabajo sindical en ese aspecto. De la misma manera, las áreas donde la mayoría o la totalidad sí se encuentra informada, demuestran entonces un buen trabajo sindical y administrativo.

5.10. Prueba X^2 para determinar concordancia casual entre expertos.

Véase un ejemplo.

Ejemplo 8:

Un estudiante de la carrera de Licenciatura en Turismo, identificó 11 actividades llevadas a cabo en el sub-proceso de check-in, correspondiente al proceso de Recepción, del área de alojamiento en el Hotel X.

Estas actividades las sometió a criterio de los expertos del hotel para que las ordenaran según el orden de prioridad, que a su juicio, debían tener.

El criterio para ordenar, fue el grado de utilidad que cada una de esas actividades, le reportaba a los clientes. El resultado de este trabajo fue el siguiente:

Sub-proceso	No.	Actividades	EXPERTOS						
			E_1	E_2	E_3	E_4	E_5	E_6	E_7
Check-in	1	Recibir clientes en la entrada	1	3	2	2	1	1	1
	2	Realizar cóctel de bienvenida	5	5	3	4	4	3	4
	3	Saludar al cliente en la recepción	3	2	5	1	3	8	3
	4	Solicitar la reserva	2	1	1	3	2	2	2
	5	Verificar si la reserva está hecha	4	6	7	5	5	6	5
	6	Entregar al cliente formularios	11	11	10	10	10	10	11
	7	Llenar formularios (el cliente)	7	8	6	8	9	5	8
	8	Recibir formularios del cliente	9	9	8	9	8	7	9
	9	Asignar la habitación	8	7	9	7	7	9	7
	10	Entregar la llave de la habitación	6	4	4	6	6	4	6
	11	Conducir al cliente a la habitación	10	10	11	11	11	11	10

El estudiante desea saber, si existe o no concordancia entre los juicios emitidos por los expertos a la hora de ordenar cada actividad, por orden de importancia o utilidad para el cliente.

En caso de existir concordancia, deberá proceder a verificar si la misma es o no casual, con un nivel de confiabilidad del 95%.

Solución:

El primer paso consiste en determinar, si existe o no concordancia entre los juicios emitidos por los expertos. Antes de trabajar con el SPSS, nótese que la tabla de datos anterior, contiene la cantidad de expertos (decisores) colocada en forma de columnas (7 columnas) y la cantidad de criterios a ordenar por importancia (actividades del sub-proceso de check-in) en forma de filas (11 filas).

De este modo, no es posible colocar los datos en el SPSS, pues debe recordarse que SIEMPRE es preciso que las variables o atributos a evaluar, sean colocados en forma de columnas. Para ello se requiere transponer la matriz de datos, de modo que las 11 actividades de check-in (atributos) aparezcan en 11 columnas. (La transposición de una matriz se realiza fácilmente mediante el Microsoft Excel).

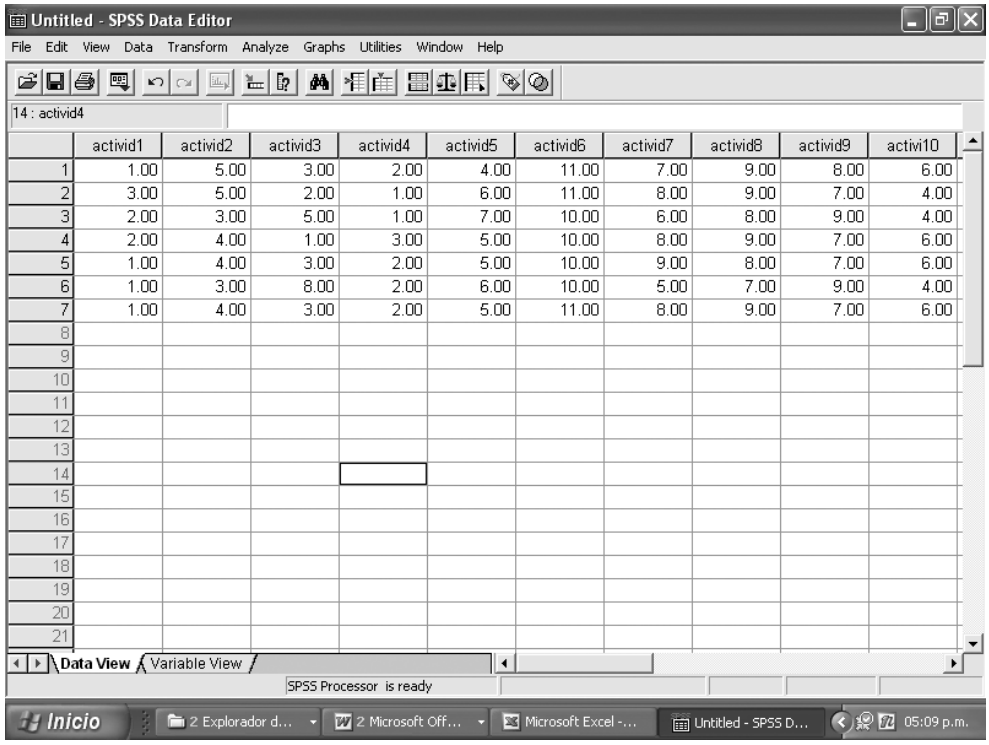
Matriz de datos original (Excel)

	E_1	E_2	E_3	E_4	E_5	E_6	E_7
A_1	1	3	2	2	1	1	1
A_2	5	5	3	4	4	3	4
A_3	3	2	5	1	3	8	3
A_4	2	1	1	3	2	2	2
A_5	4	6	7	5	5	6	5
A_6	11	11	10	10	10	10	11
A_7	7	8	6	8	9	5	8
A_8	9	9	8	9	8	7	9
A_9	8	7	9	7	7	9	7
A_{10}	6	4	4	6	6	4	6
A_{11}	10	10	11	11	11	11	10

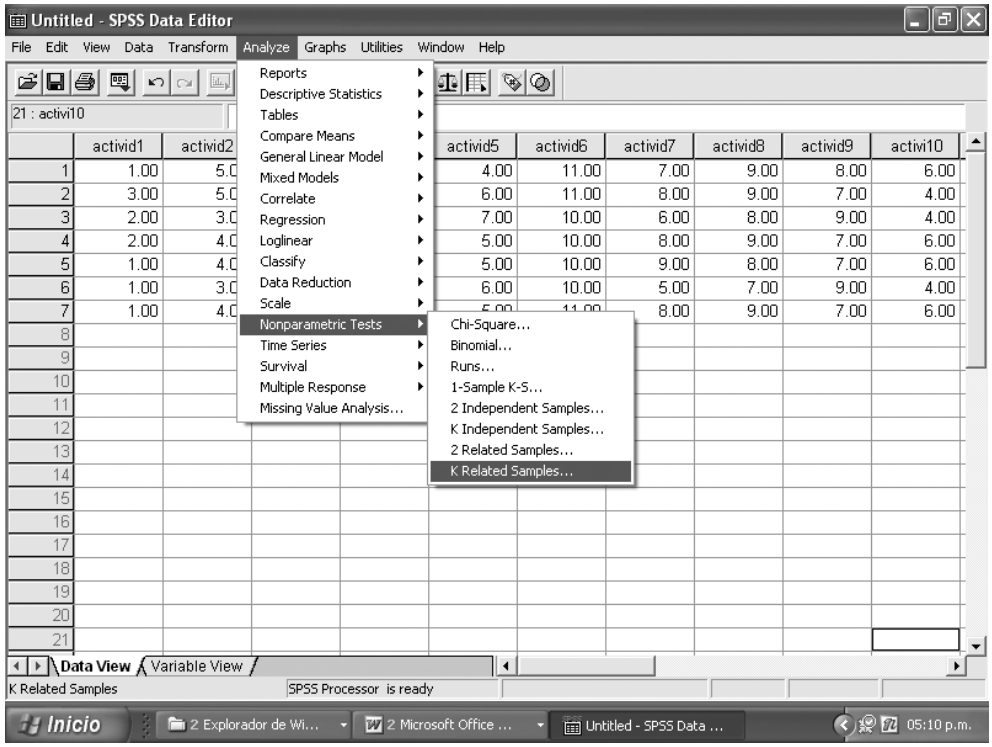
Matriz de datos transpuesta (Excel)

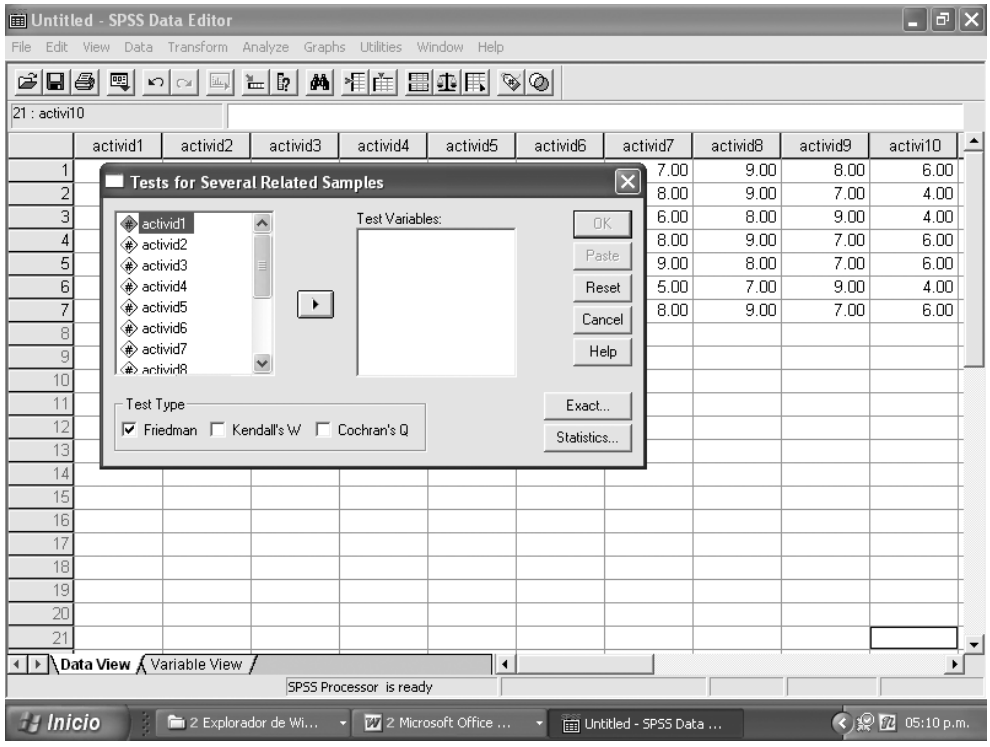
	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}
E_1	1	5	3	2	4	11	7	9	8	6	10
E_2	3	5	2	1	6	11	8	9	7	4	10
E_3	2	3	5	1	7	10	6	8	9	4	11
E_4	2	4	1	3	5	10	8	9	7	6	11
E_5	1	4	3	2	5	10	9	8	7	6	11
E_6	1	3	8	2	6	10	5	7	9	4	11
E_7	1	4	3	2	5	11	8	9	7	6	10

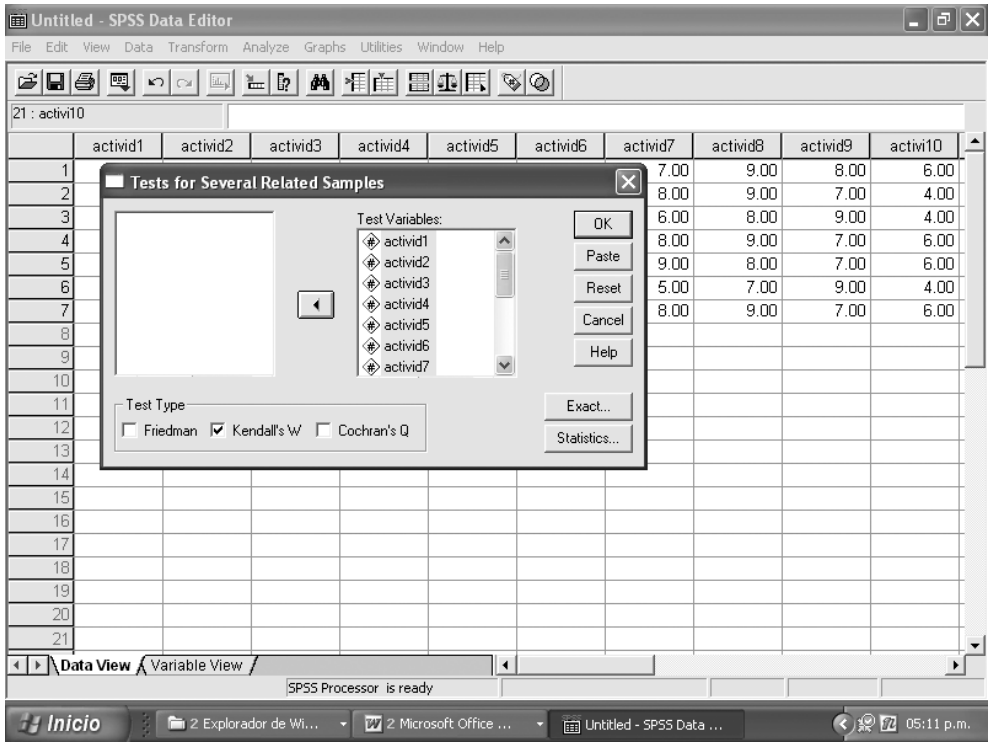
Ahora estando listos los datos para ser colocados en el SPSS, sería:

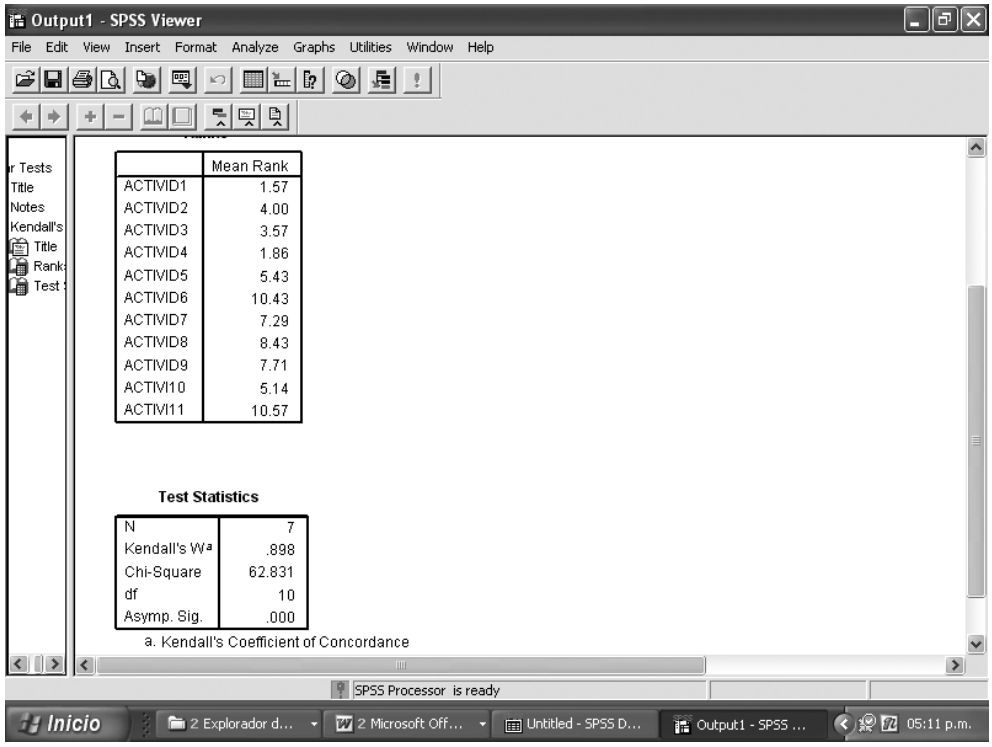


	actividad1	actividad2	actividad3	actividad4	actividad5	actividad6	actividad7	actividad8	actividad9	actividad10
1	1.00	5.00	3.00	2.00	4.00	11.00	7.00	9.00	8.00	6.00
2	3.00	5.00	2.00	1.00	6.00	11.00	8.00	9.00	7.00	4.00
3	2.00	3.00	5.00	1.00	7.00	10.00	6.00	8.00	9.00	4.00
4	2.00	4.00	1.00	3.00	5.00	10.00	8.00	9.00	7.00	6.00
5	1.00	4.00	3.00	2.00	5.00	10.00	9.00	8.00	7.00	6.00
6	1.00	3.00	8.00	2.00	6.00	10.00	5.00	7.00	9.00	4.00
7	1.00	4.00	3.00	2.00	5.00	11.00	8.00	9.00	7.00	6.00




Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
21 : activi10
actividad1 actividad2 actividad3 actividad4 actividad5 actividad6 actividad7 actividad8 actividad9 activi10
Tests for Several Related Samples
actividad1
actividad2
actividad3
actividad4
actividad5
actividad6
actividad7
Test Variables:
OK
Paste
Reset
Cancel
Help
Exact...
Statistics...
Test Type
Friedman
Kendall's W
Cochran's Q
Data View
Variable View
SPSS Processor is ready
Inicio
2 Explorador de Wi...
2 Microsoft Office ...
Untitled - SPSS Data ...
05:10 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
21 : activi10
actividad1 actividad2 actividad3 actividad4 actividad5 actividad6 actividad7 actividad8 actividad9 activi10
Tests for Several Related Samples
Test Variables:
actividad1
actividad2
actividad3
actividad4
actividad5
actividad6
actividad7
OK
Paste
Reset
Cancel
Help
Exact...
Statistics...
Test Type
Friedman
Kendall's W
Cochran's Q
Data View
Variable View
SPSS Processor is ready
Inicio
2 Explorador de Wi...
2 Microsoft Office ...
Untitled - SPSS Data ...
05:11 p.m.



El coeficiente de concordancia de Kendall (W) es igual a 0.898 indicando que los siete expertos, efectivamente concuerdan, pues el valor de W se halla entre 0.5 y 1 (coinciden en sus juicios en un 89,8%).

Habiendo existido concordancia entre los especialistas del tema, el segundo paso consiste en verificar si dicha coincidencia es o no casual, con un nivel de significación del 5%. Aquí es donde se lleva a cabo la prueba de hipótesis no paramétrica X^2 que da nombre a este epígrafe.

H_0: concordancia casual (los expertos coinciden por azar)

H_1: concordancia no casual (los expertos coinciden no por casualidad, sino por su conocimiento acerca del tema)

Obsérvese que debajo del valor del coeficiente de concordancia de Kendall (W), aparece el valor del estadígrafo X^2 de la dócima antes mencionada. Como el valor de probabilidad de la prueba es igual a 0.000 y dicho valor es menor que 0.05, entonces se cumple la región crítica y se rechaza la hipótesis nula. Esto indica que la concordancia entre los juicios emitidos por los expertos, es no casual. De esta forma pudiera afirmarse también que, sin dudas, los siete sujetos tomados para el estudio, son expertos.

Nota: vale la pena señalar, que el Método del Coeficiente de Concordancia de Kendall constituye, además, un Método de Expertos, o sea, que sirve para probar experticidad de un grupo de sujetos, tomados como posibles expertos en un tema.

EJERCITACIÓN

Un grupo de trabajo de la Transportista H, ha registrado el promedio diario de kilómetros recorridos por cada uno de sus 16 ómnibus nuevos que fueron adquiridos hace un mes. Esos buses se emplean para realizar los circuitos por todo el país, que las agencias de viajes del polo venden a los turistas, por tanto, son equipos sometidos, en poco tiempo, a un kilometraje elevado. El grupo de trabajo desea saber si los kilómetros recorridos por cada ómnibus hasta la fecha, siguen o no una distribución normal con un nivel de significación del 5%. Los datos son los siguientes:

Ómnibus	Promedio de kilómetros diarios
1	2344.53
2	1674.89
3	3109.80
4	2564.98
5	989.68
6	1657.57
7	3527.60
8	2659.36
9	1756.37
10	2758.98
11	2768.45
12	967.65
13	2657.58
14	1649.35
15	3546.10
16	3817.67

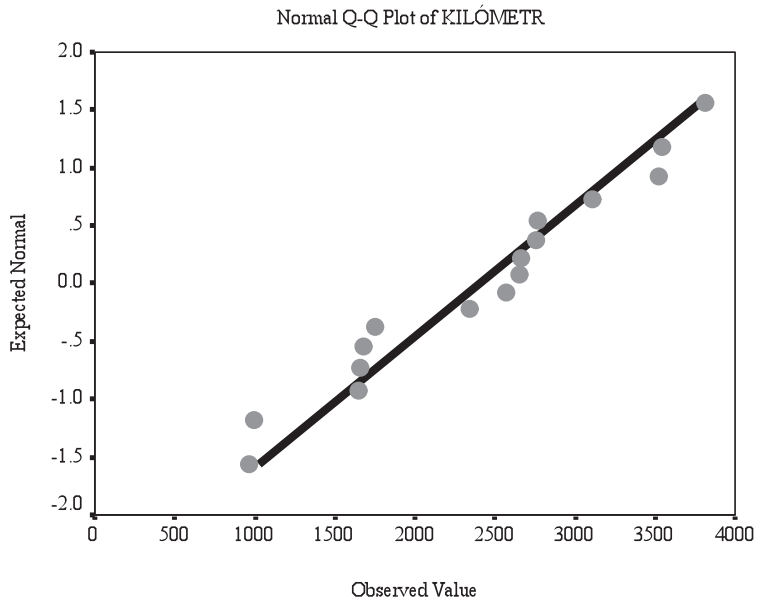
SOLUCIÓN

H_0: $X \sim N(\mu; \sigma^2)$ (la cantidad de kilómetros diarios recorridos, sigue una distribución normal)

H_1: $X \overset{no}{\sim} N(\mu; \sigma^2)$ (la cantidad de kilómetros diarios recorridos, no sigue una distribución normal)

Valor de probabilidad de la dócima: 0.437

El grupo de trabajo de la Transportista H, ha podido comprobar que bajo un nivel de confiabilidad del 95%, la cantidad promedio de kilómetros diarios recorridos por los ómnibus, sí sigue una distribución normal. Esto indica que la explotación de todos esos buses de nueva adquisición, está siendo pareja, de modo que ninguno está siendo sobreutilizado respecto a los demás. El gráfico también muestra esta conclusión:



Capítulo
6

Análisis de varianza.

6.1. Generalidades acerca del análisis de varianza (ANOVA).

El análisis de varianza constituye una dócima de hipótesis paramétrica.

Como prueba estadística, se emplea para analizar si más de dos grupos difieren significativamente entre sí en cuanto a sus medias y sus varianzas.

6.2. Requisitos para llevar a cabo un análisis de varianza.

Son tres los supuestos sobre los cuales descansa este análisis:

- las observaciones de cada nivel del factor, deben ser tomadas aleatoriamente y de forma independiente (puede garantizarse mediante un muestreo aleatorio simple)
- las observaciones, de cada nivel del factor, deben seguir una distribución normal (puede verificarse mediante una prueba de bondad de ajuste)
- las varianzas de todos los niveles del factor, deben ser homogéneas (homocedasticidad) (esto puede verificarse mediante la prueba de hipótesis de Cochran o la de Bartlett, según corresponda)

Véase un ejemplo de análisis de varianza con un solo factor (ANOVA one-way).

Ejemplo 1:

La Agencia de Viajes M ubicada en el polo turístico Jardines del Rey, ha informado recientemente que los niveles de venta del producto “Guamá”,

están variando de acuerdo a las nacionalidades de los clientes que compran dicha excursión u opcional.

El Departamento de Opcionales ha observado que los franceses, optan más por comprar el producto "Guamá" que los canadienses, británicos y argentinos. Por ello, la dirección del departamento ha decidido llevar a cabo una fuerte campaña promocional de este opcional, dirigida a los potenciales clientes de Canadá, Gran Bretaña y Argentina.

Después de tres meses de campaña ininterrumpida, la agencia desea conocer si esta última, ha surtido efecto en los niveles de venta de la excusión "Guamá", y ha comenzado un estudio durante cinco meses con un nivel de confiabilidad del 99%. Pasado ese tiempo, los datos son los que muestran a continuación:

Países	**Ventas / mes** (en miles de CUC)				
Canadá	200	15	34	150	28
Francia	78	102	320	17	26
Gran Bretaña	89	54	304	180	58
Argentina	20	79	36	400	174

Solución:

Variable aleatoria observada (vao): nivel de ventas

Factor: países

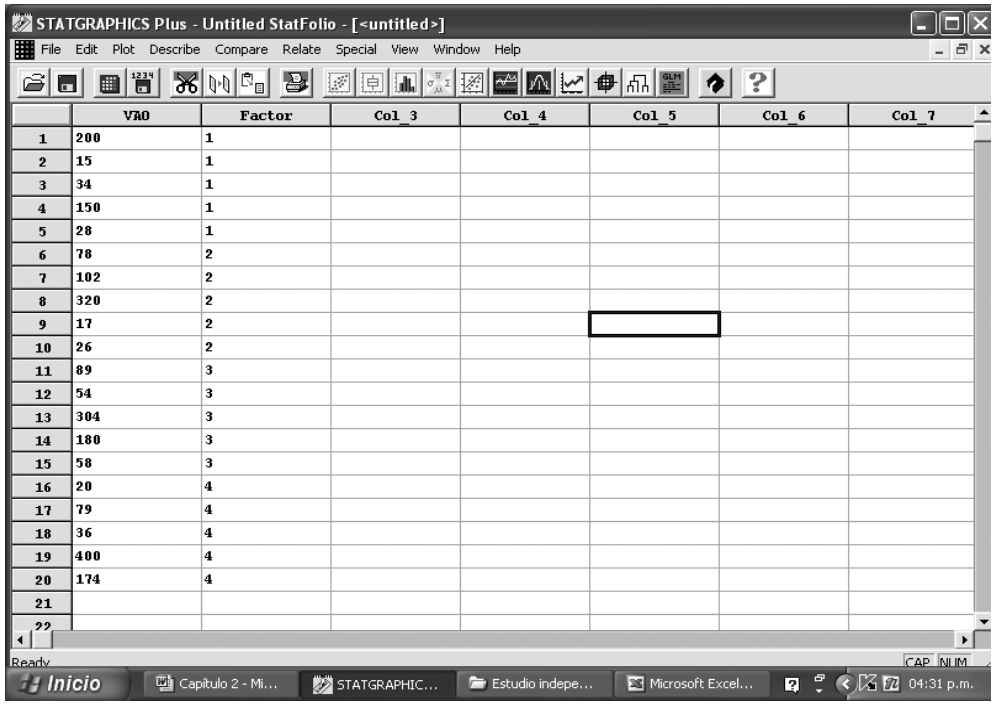
Niveles del factor: 4

Cantidad de observaciones por cada nivel del factor: 5

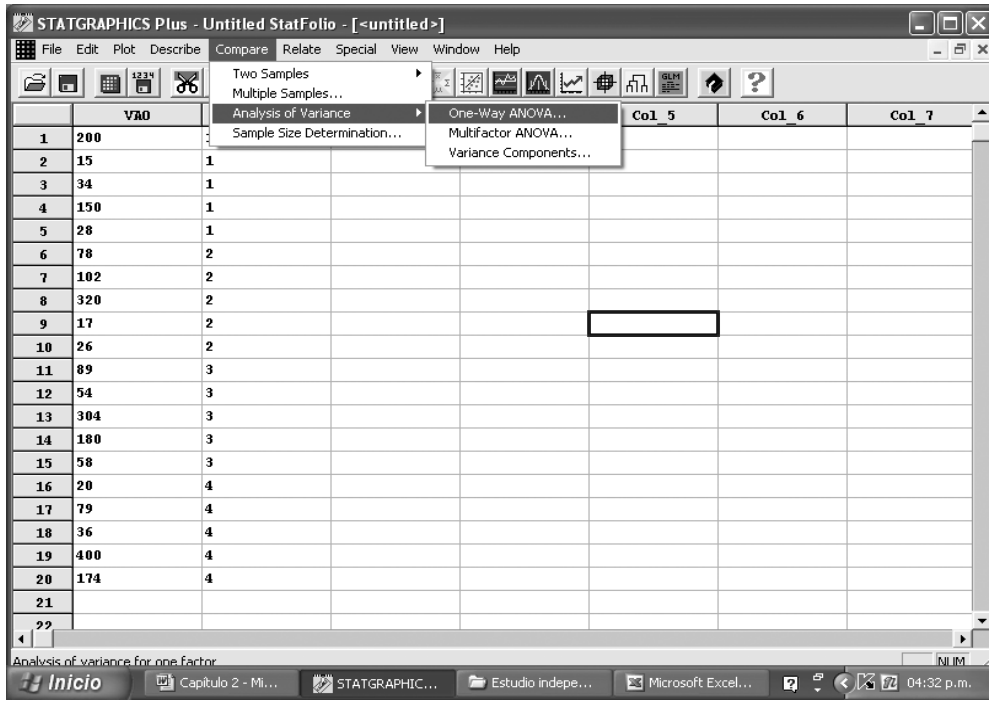
H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$ (la nacionalidad no influye en el nivel de venta)

H_1: algún $\mu_j \neq \mu$ (la nacionalidad sí influye en el nivel de venta)

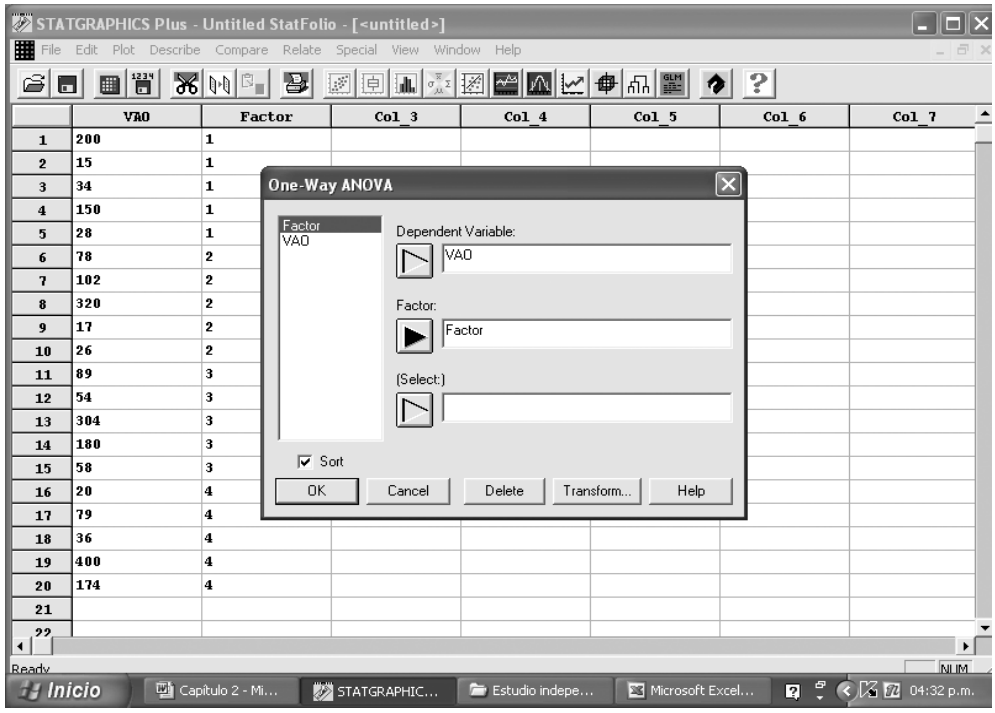
Empleando el STATGRAPHICS para procesar los datos y realizar la dócima, sería:

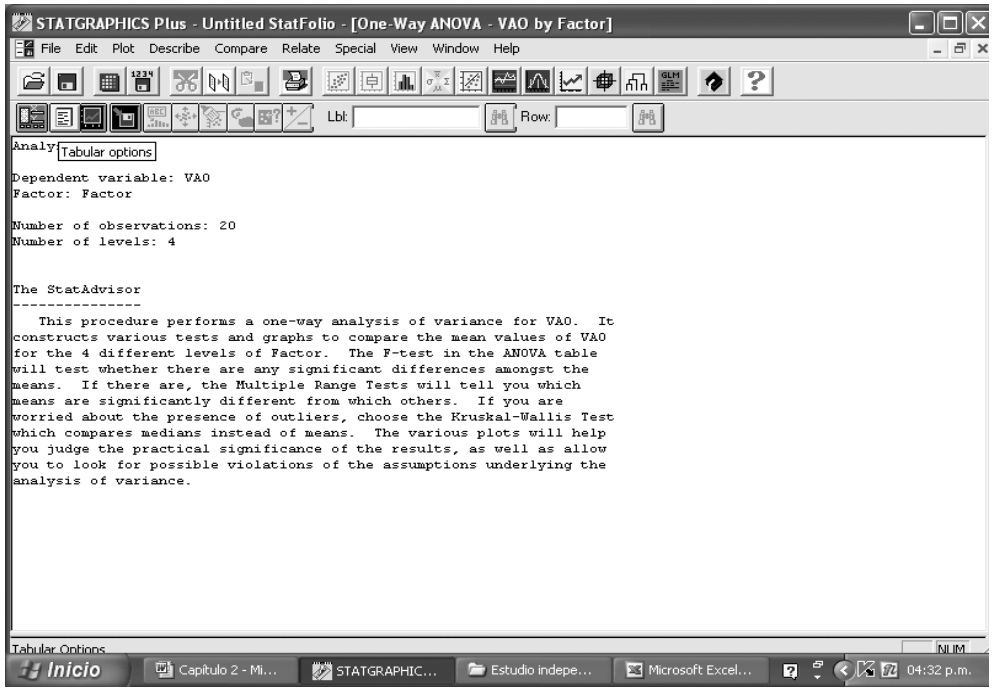


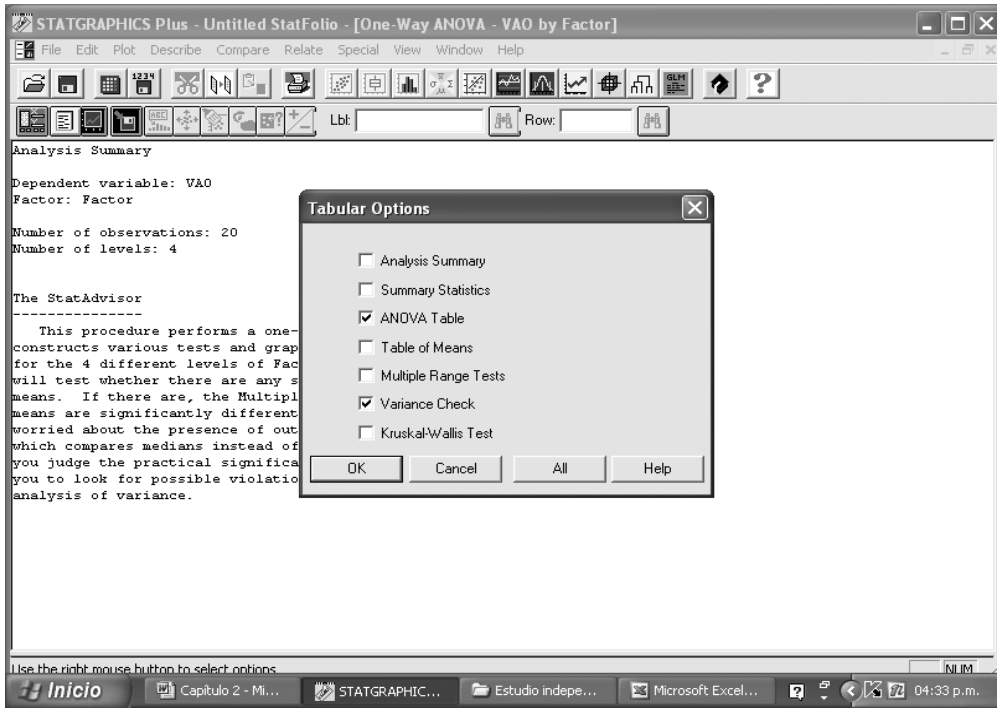
	VAO	Factor	Col_3	Col_4	Col_5	Col_6	Col_7
1	200	1					
2	15	1					
3	34	1					
4	150	1					
5	28	1					
6	78	2					
7	102	2					
8	320	2					
9	17	2					
10	26	2					
11	89	3					
12	54	3					
13	304	3					
14	180	3					
15	58	3					
16	20	4					
17	79	4					
18	36	4					
19	400	4					
20	174	4					
21							
22							

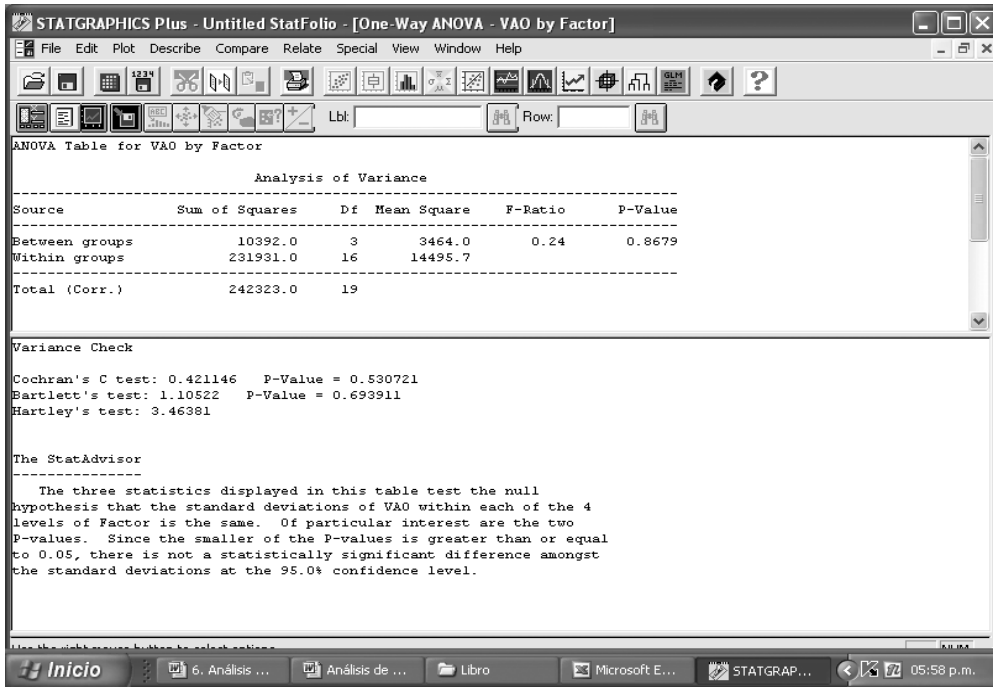


	VAO	Col_5	Col_6	Col_7
1	200			
2	15			
3	34			
4	150			
5	28			
6	78			
7	102			
8	320			
9	17			
10	26			
11	89			
12	54			
13	304			
14	180			
15	58			
16	20			
17	79			
18	36			
19	400			
20	174			
21				
22				


STATGRAPHICS Plus - Untitled StatFolio - [<untitled>]
File Edit Plot Describe Compare Relate Special View Window Help
VAO
Factor
Col_3
Col_4
Col_5
Col_6
Col_7
1 200 1
2 15 1
3 34 1
4 150 1
5 28 1
6 78 2
7 102 2
8 320 2
9 17 2
10 26 2
11 89 3
12 54 3
13 304 3
14 180 3
15 58 3
16 20 4
17 79 4
18 36 4
19 400 4
20 174 4
21
22
One-Way ANOVA
Factor
VAO
Dependent Variable:
VAO
Factor:
Factor
(Select:)
Sort
OK
Cancel
Delete
Transform...
Help
Ready
NUM
Inicio
Capitulo 2 - Mi...
STATGRAPHIC...
Estudio indepe...
Microsoft Excel...
04:32 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [One-Way ANOVA - VAO by Factor]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl:
Row:
Analy
Tabular options
Dependent variable: VAO
Factor: Factor
Number of observations: 20
Number of levels: 4
The StatAdvisor
This procedure performs a one-way analysis of variance for VAO. It constructs various tests and graphs to compare the mean values of VAO for the 4 different levels of Factor. The F-test in the ANOVA table will test whether there are any significant differences amongst the means. If there are, the Multiple Range Tests will tell you which means are significantly different from which others. If you are worried about the presence of outliers, choose the Kruskal-Wallis Test which compares medians instead of means. The various plots will help you judge the practical significance of the results, as well as allow you to look for possible violations of the assumptions underlying the analysis of variance.
Tabular Options
NUM
Inicio
Capitulo 2 - Mi...
STATGRAPHIC...
Estudio indepe...
Microsoft Excel...
04:32 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [One-Way ANOVA - VAO by Factor]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
Analysis Summary
Dependent variable: VAO
Factor: Factor
Number of observations: 20
Number of levels: 4
The StatAdvisor
This procedure performs a one-
constructs various tests and grap
for the 4 different levels of Fac
will test whether there are any s
means. If there are, the Multipl
means are significantly different
worried about the presence of out
which compares medians instead of
you judge the practical significa
you to look for possible violatio
analysis of variance.
Tabular Options
Analysis Summary
Summary Statistics
ANOVA Table
Table of Means
Multiple Range Tests
Variance Check
Kruskal-Wallis Test
OK
Cancel
All
Help
Use the right mouse button to select options
Inicio
Capitulo 2 - Mi...
STATGRAPHIC...
Estudio indepe...
Microsoft Excel...
04:33 p.m.


STATGRAPHICS Plus - Untitled StatFolio - [One-Way ANOVA - VAO by Factor]
File Edit Plot Describe Compare Relate Special View Window Help
Lbl: Row:
ANOVA Table for VAO by Factor
Analysis of Variance
Source Sum of Squares Df Mean Square F-Ratio P-Value
Between groups 10392.0 3 3464.0 0.24 0.8679
Within groups 231931.0 16 14495.7
Total (Corr.) 242323.0 19
Variance Check
Cochran's C test: 0.421146 P-Value = 0.530721
Bartlett's test: 1.10522 P-Value = 0.693911
Hartley's test: 3.46381
The StatAdvisor
The three statistics displayed in this table test the null
hypothesis that the standard deviations of VAO within each of the 4
levels of Factor is the same. Of particular interest are the two
P-values. Since the smaller of the P-values is greater than or equal
to 0.05, there is not a statistically significant difference amongst
the standard deviations at the 95.0% confidence level.
Inicio
6. Análisis ...
Análisis de ...
Libro
Microsoft E...
STATGRAP...
05:58 p.m.

Suponiendo que se cumplen los dos primeros requisitos mencionados en el epígrafe 6.2, obsérvese que se ha escogido la opción de realizar también la prueba que demuestra si existe o no homocedasticidad. La dócima de Cochran se lleva a cabo con modelos equilibrados, o sea, cuando la cantidad de observaciones tomadas por cada nivel del factor, es la misma. La de Bartlett, por el contrario, se realiza cuando los modelos no son equilibrados (se toman diferentes cantidades de observaciones por cada nivel del factor). El ejemplo representado constituye un modelo equilibrado de análisis de varianza, pues por cada uno de los 4 niveles en que se divide el factor (países), se toman 5 observaciones de la variable (nivel de venta).

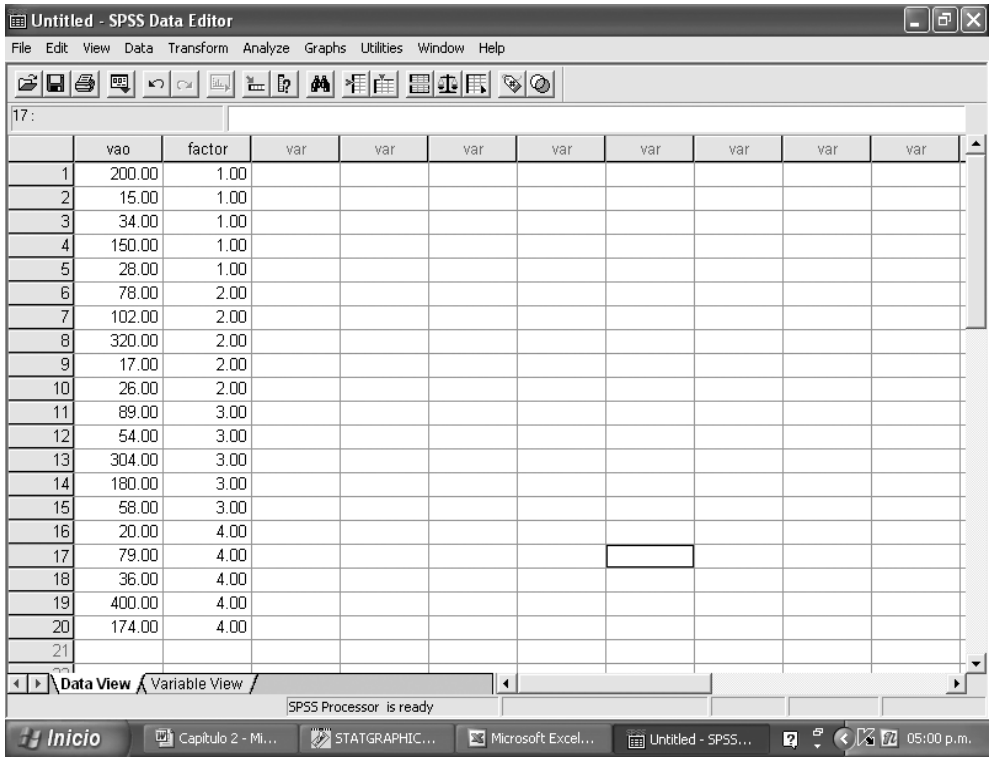
Dócima de Cochran:

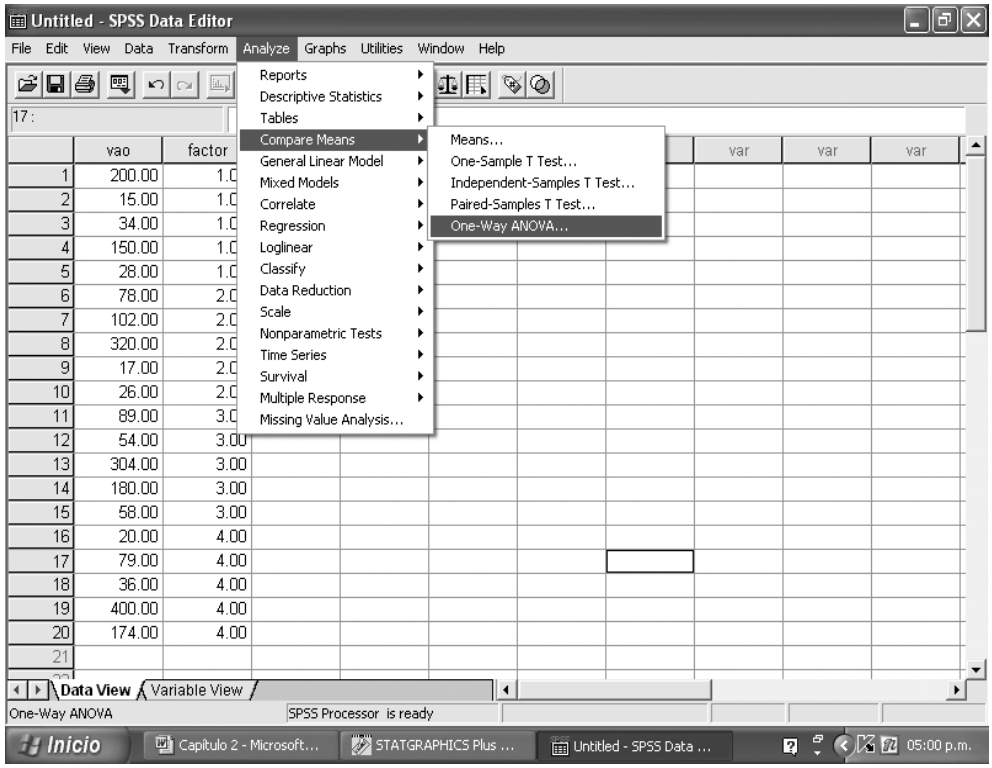
H_0: $\sigma^2_1 = \sigma^2_2 = \sigma^2_3 = \sigma^2_4 = \sigma^2$ (existe homogeneidad de varianzas)
H_1: algún $\sigma^2_j \neq \sigma^2$ (no existe homogeneidad de varianzas)

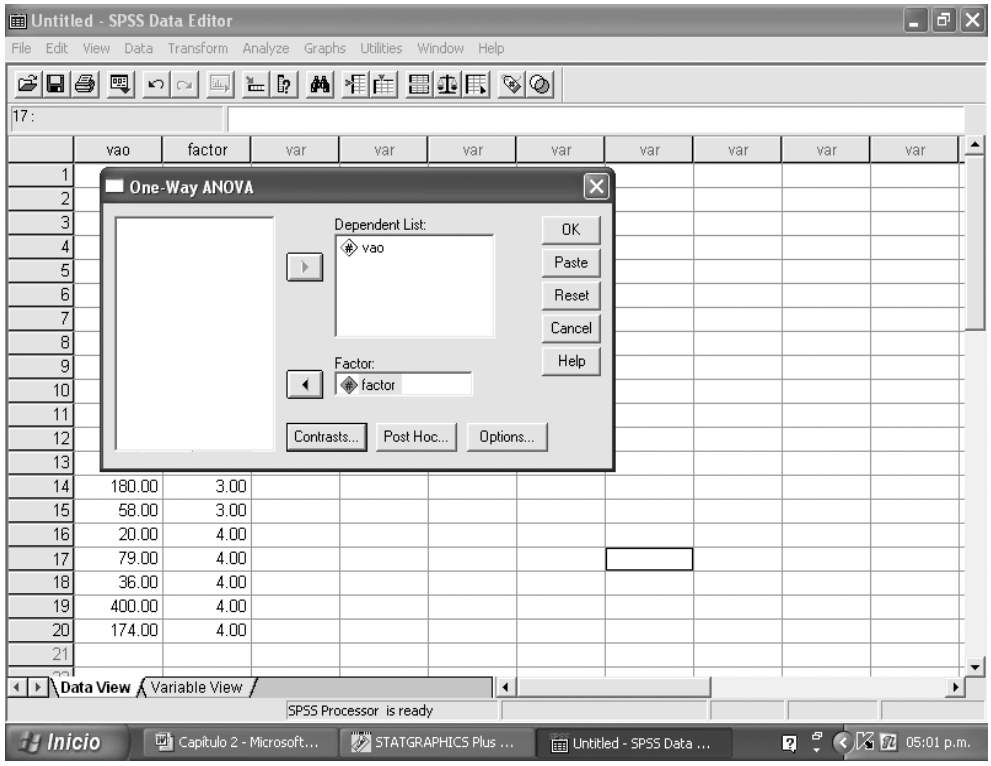
Como en la prueba de Cochran el valor de probabilidad es igual a 0.53 y el mismo no es menor que 0.01, entonces no se cumple la región crítica y se acepta la hipótesis nula. Esto indica que las varianzas son similares y por tanto, se cumple el supuesto de homocedasticidad.

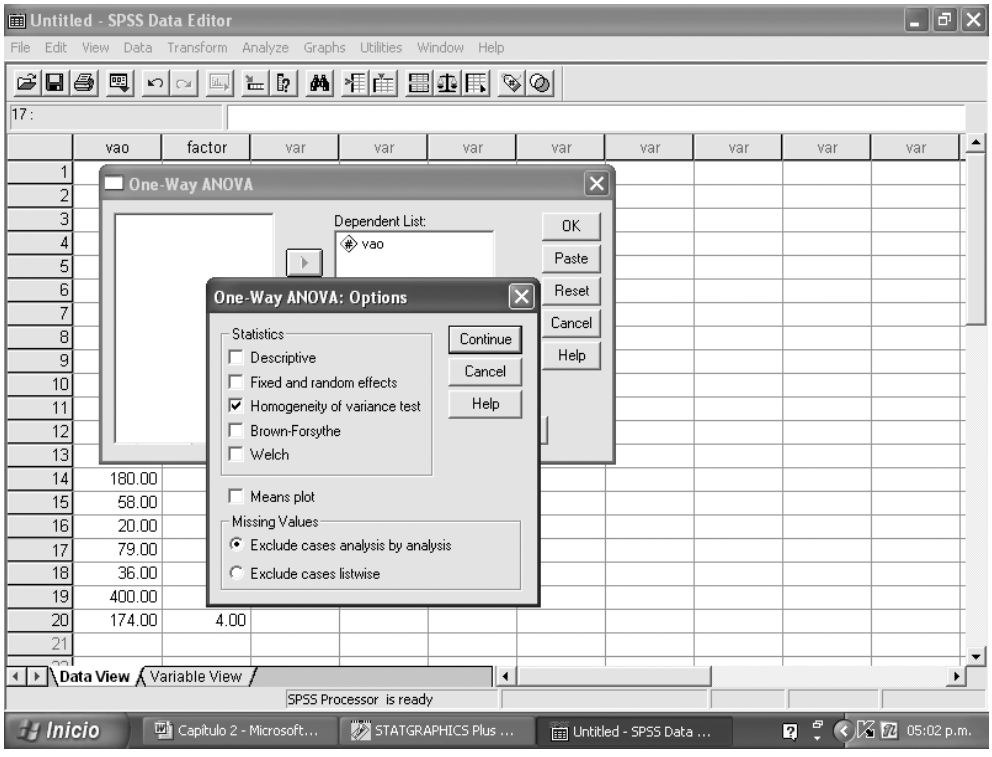
Después de comprobado el cumplimiento de este requisito o supuesto, vemos que el valor de probabilidad en la tabla ANVA (análisis de varianza) es igual a 0.87, el cual no es menor que 0.01 por lo que no se cumple la región crítica y se acepta la hipótesis nula. Esto quiere decir que el promedio de los niveles de venta por cada una de las nacionalidades de clientes (o países), ha sido similar, demostrando que el Departamento de Opcionales de la Agencia de Viajes M, puede afirmar que la campaña promocional no ha surtido el efecto deseado, con un nivel de significación del 1%.

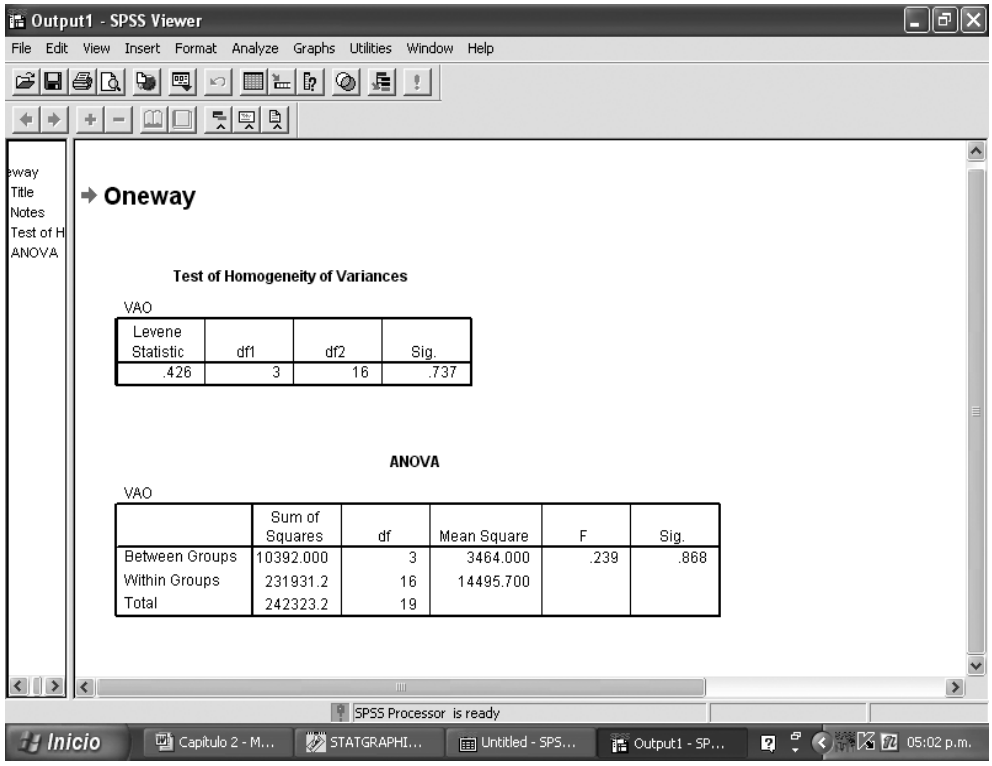
Un análisis de varianza también puede desarrollarse fácilmente utilizando el SPSS. Obsérvese cómo es:


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
17 :
vao factor var var var var var var var var
1 200.00 1.00
2 15.00 1.00
3 34.00 1.00
4 150.00 1.00
5 28.00 1.00
6 78.00 2.00
7 102.00 2.00
8 320.00 2.00
9 17.00 2.00
10 26.00 2.00
11 89.00 3.00
12 54.00 3.00
13 304.00 3.00
14 180.00 3.00
15 58.00 3.00
16 20.00 4.00
17 79.00 4.00
18 36.00 4.00
19 400.00 4.00
20 174.00 4.00
21
Data View Variable View
SPSS Processor is ready
Inicio
Capítulo 2 - Mi...
STATGRAPHIC...
Microsoft Excel...
Untitled - SPSS...
05:00 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Means...
One-Sample T Test...
Independent-Samples T Test...
Paired-Samples T Test...
One-Way ANOVA...
17 :
vao factor var var var
Data View Variable View
One-Way ANOVA
SPSS Processor is ready
Inicio
Capítulo 2 - Microsoft...
STATGRAPHICS Plus ...
Untitled - SPSS Data ...
05:00 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
One-Way ANOVA
Dependent List:
vao
Factor:
factor
OK
Paste
Reset
Cancel
Help
Contrasts...
Post Hoc...
Options...
Data View
Variable View
SPSS Processor is ready
Inicio
Capítulo 2 - Microsoft...
STATGRAPHICS Plus ...
Untitled - SPSS Data ...
05:01 p.m.


Untitled - SPSS Data Editor
One-Way ANOVA
One-Way ANOVA: Options
Statistics
Descriptive
Fixed and random effects
Homogeneity of variance test
Brown-Forsythe
Welch
Means plot
Missing Values
Exclude cases analysis by analysis
Exclude cases listwise
Continue
Cancel
Help
Data View
Variable View
SPSS Processor is ready
05:02 p.m.



EJERCITACIÓN

Un grupo de estudiantes de la carrera de Licenciatura en Turismo, ha decidido comprobar si la demora en la atención a los clientes durante la actividad de check-out en un hotel, depende de la categoría o estrellaje del mismo. Para ello, han visitado tres hoteles de categorías 3, 4 y 5 estrellas respectivamente, y han tomado de cada uno, el promedio diario, durante seis días, de demora de la actividad de check-out en el área de Recepción, con un nivel de confiabilidad del 99%. Los datos se hallan a continuación:

Hotel	**Tiempo promedio** (minutos)					
Villa Playa Azul (3 estrellas)	3	4	7	5	4	6
Sol Caimán Verde (4 estrellas)	2	4	1	2	2	3
Tryp Palma Real (5 estrellas)	1	3	1	1	1	2

SOLUCIÓN

H_0: $\sigma^2_1 = \sigma^2_2 = \sigma^2_3 = \sigma^2$

H_1: algún $\sigma^2_j \neq \sigma^2$

Valor de probabilidad de la dócima de Cochran: 0.331

Se cumple el supuesto de homocedasticidad.

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu$

H_1: algún $\mu_j \neq \mu$

Valor de probabilidad de la dócima: 0.000

El grupo de estudiantes ha podido comprobar, que la demora en la atención a los clientes durante la actividad de check-out, sí depende de la categoría del hotel, con un nivel de significación del 1%. Este era el resultado lógico que esperaban obtener los estudiantes, pues es de suponer, por ejemplo, que en un hotel con categoría de 5 estrellas, la demora de cualquier servicio sea mucho menor que la de otros hoteles con menor categoría.

Capítulo
7

Análisis de asociación.

7.1. Generalidades acerca de las distribuciones bidimensionales.

Estas distribuciones aparecen cuando de una población se desean estudiar dos variables.

Existen varios tipos de distribuciones bidimensionales:
- cuando las dos informaciones son atributos
- cuando una información corresponde a una variable y la otra a un atributo
- cuando las dos informaciones son variables

Cuando las variables son cuantitativas, a las tablas de frecuencias se les denomina "tablas de correlación", y cuando se trata de atributos o variables cualitativas, se les llama "tablas de contingencia".

Nótese que en las distribuciones bidimensionales, lo que más se realiza es análisis de correlación, o sea, determinar si existe o no relación entre las dos informaciones (del tipo que sea cada una), y en caso de que sí, realizar un estudio detallado de dicha relación.

En ocasiones, ha ocurrido que cierta teoría supone la relación entre dos informaciones, pero al intentar demostrarla en la práctica, no se evidencia ninguna. A esto se le denomina: correlación espuria (aparente).

Para determinar qué correlación existe entre variables cualitativas, se emplean los siguientes coeficientes:
- coeficiente de Goodman y Kruskall
- coeficiente de Kendall
- coeficiente de Yule

- coeficiente Chi-Cuadrado
- coeficiente Phi

Para determinar qué correlación existe entre variables cuantitativas, se emplean los siguientes coeficientes:

- coeficiente de correlación por rangos de Spearman
- coeficiente de correlación de Pearson

Este libro hará énfasis en el estudio acerca de los coeficientes de correlación de Spearman y Pearson, por ser los de utilización más frecuente en el ámbito del turismo, aunque por supuesto, el resto de los coeficientes cumple cada uno su función donde no existe uno mejor que otro.

7.2. Coeficiente de correlación por rangos de Spearman.

La correlación de Spearman (rho) constituye un excelente método para cuantificar la relación entre dos escalas de valores cuantitativos discretos y/o con jerarquía (ordinales). También es una excelente opción cuando los datos de las variables no tienen una distribución normal bivariante, especialmente si hay valores extremos.

Permite determinar el grado de relación o asociación, pero sólo entre dos variables, a diferencia del coeficiente de correlación de Pearson. Su interpretación es muy similar a la de este último coeficiente, pues sus valores oscilan entre -1 y 1.

Véase un ejemplo.

Ejemplo 1:

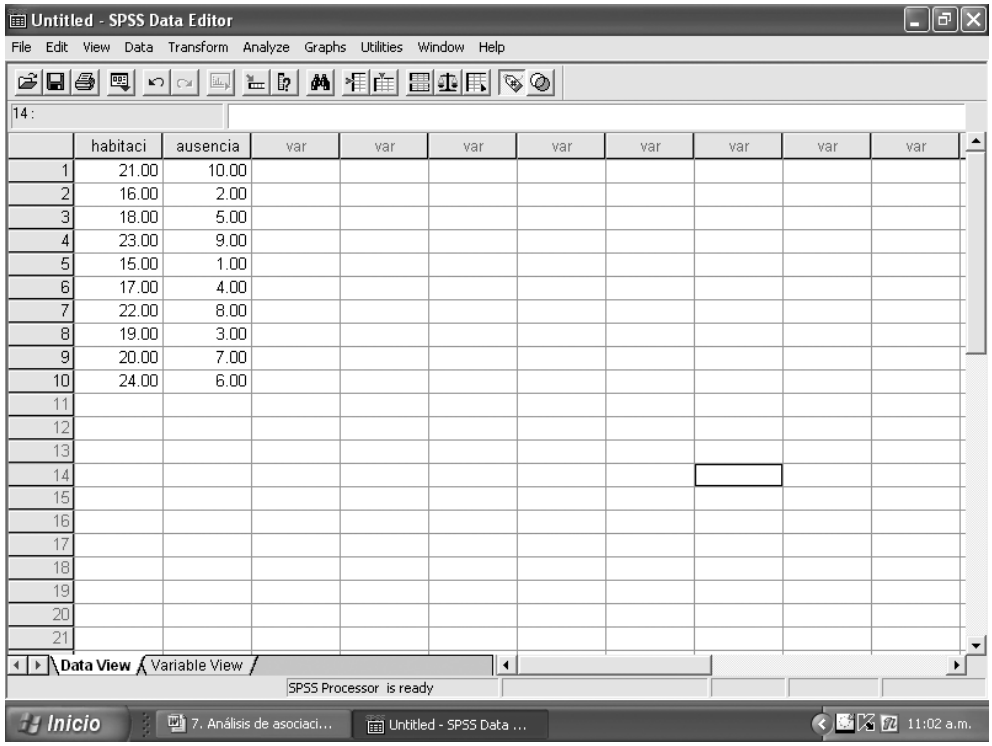
En el Hotel M ubicado en polo turístico de Cayo Guillermo, el Departamento de Recursos Humanos está realizando un estudio como parte de un Diplomado en Psicología Organizacional que están cursando sus miembros. El mismo consiste, en su primera etapa, en determinar si se cumple la sospecha de que mientras más habitaciones le asignan a las camareras (por encima de su

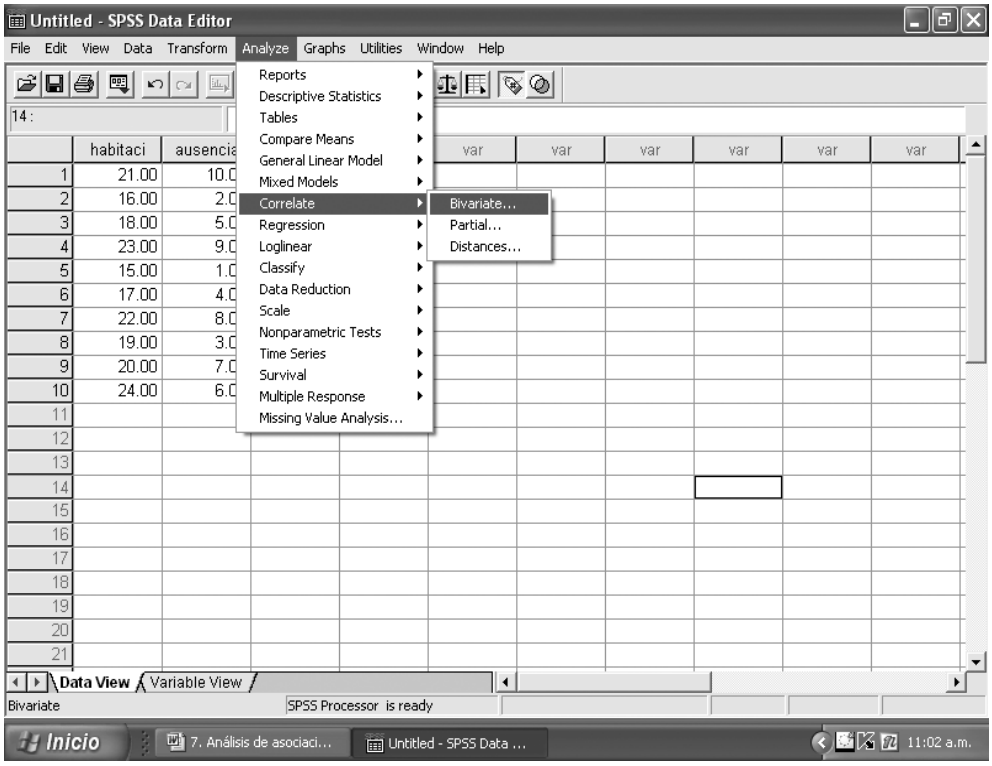
norma establecida) para la limpieza, más cantidad de ausencias al puesto de trabajo presentan las mismas, lo cual provoca deficiencias en el servicio de alojamiento brindado al cliente externo. El nivel de confiabilidad es del 99%. Los datos se muestran a continuación:

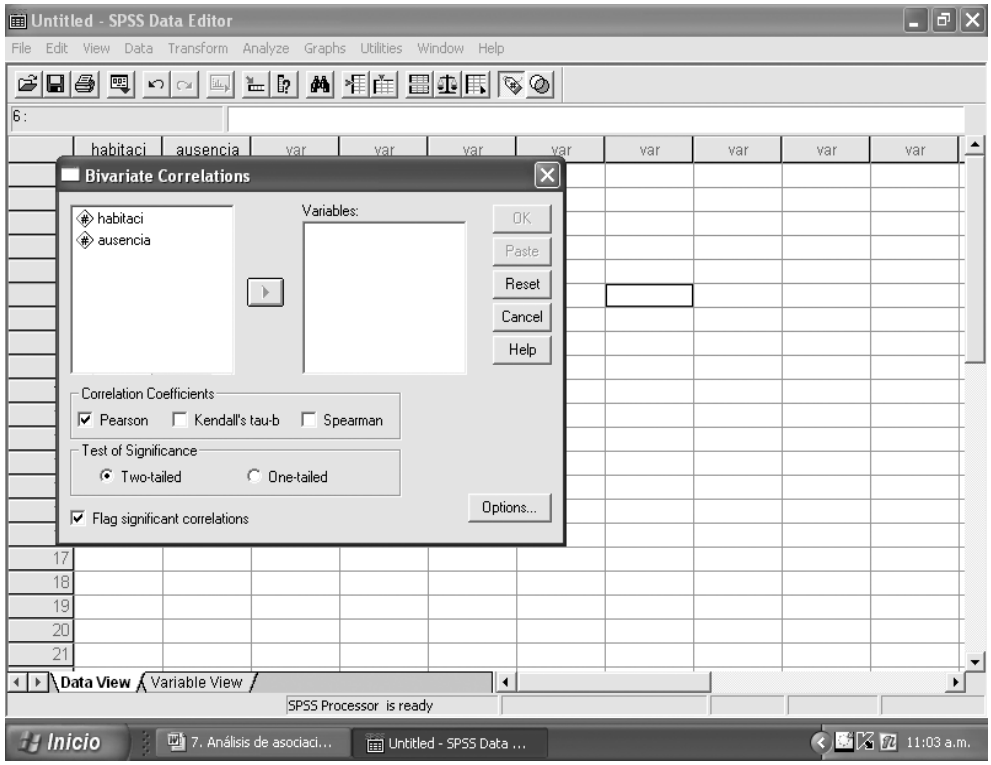
Camareras	Habitaciones asignadas	Ausencias
1	21	10
2	16	2
3	18	5
4	23	9
5	15	1
6	17	4
7	22	8
8	19	3
9	20	7
10	24	6

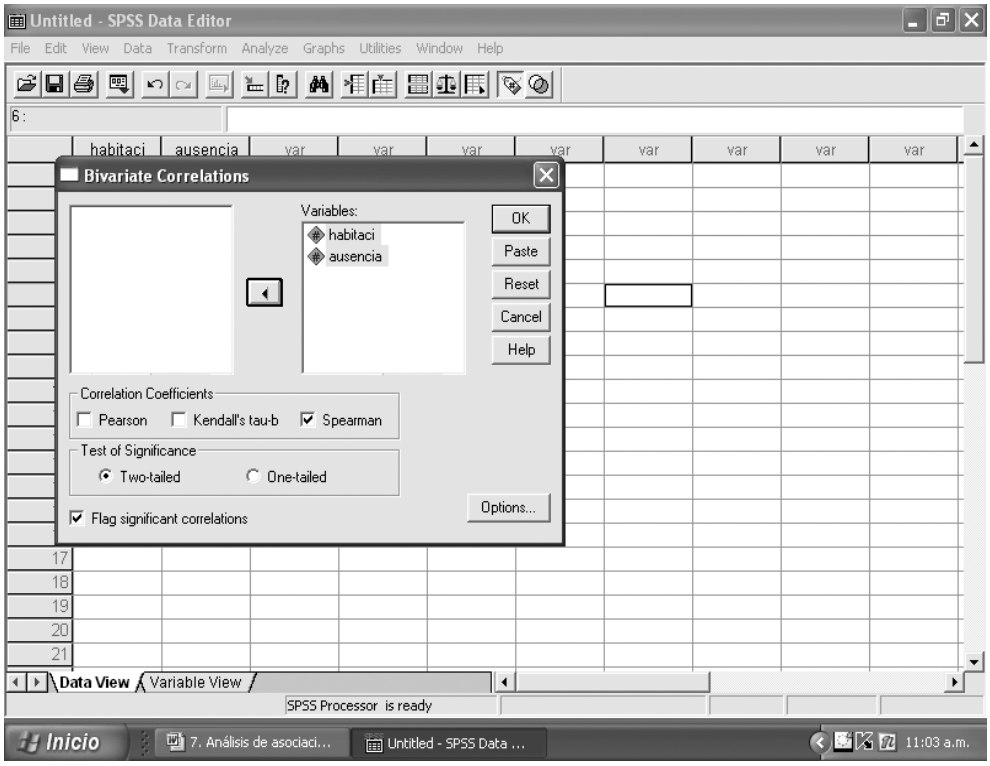
Solución:

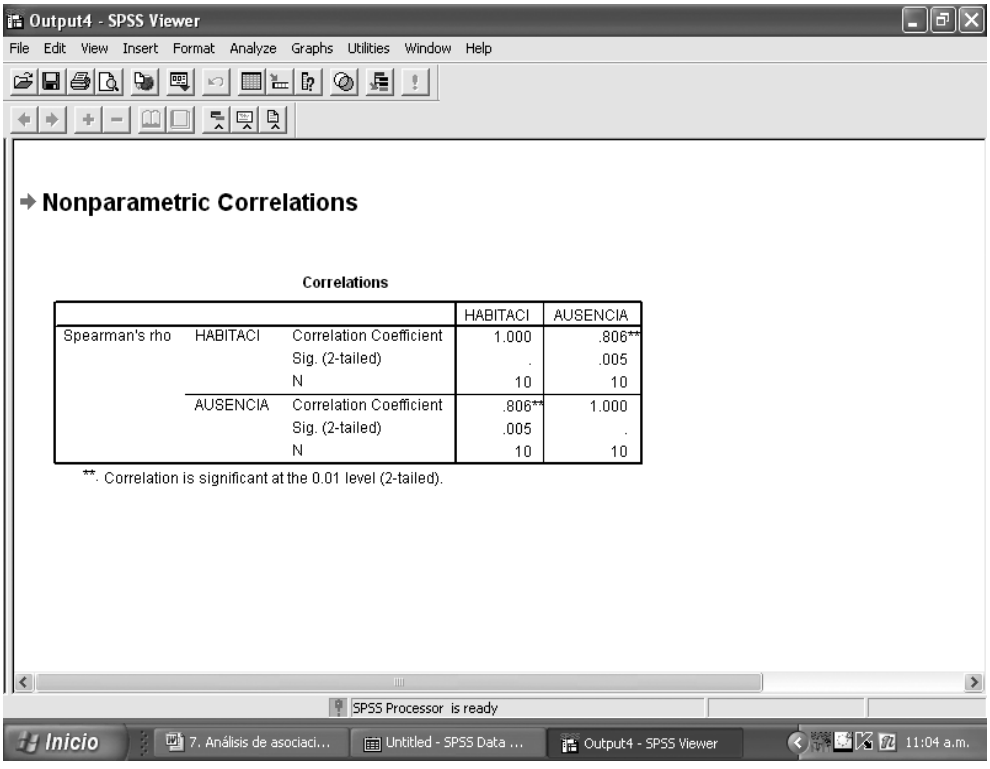
Utilizando el SPSS, sería:




Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Bivariate...
Partial...
Distances...
habitaci
ausencia
Data View
Variable View
Bivariate
SPSS Processor is ready
Inicio
7. Análisis de asociaci...
Untitled - SPSS Data ...
11:02 a.m.


Untitled - SPSS Data Editor
Bivariate Correlations
habitaci
ausencia
Variables:
OK
Paste
Reset
Cancel
Help
Correlation Coefficients
Pearson
Kendall's tau-b
Spearman
Test of Significance
Two-tailed
One-tailed
Flag significant correlations
Options...
Data View
Variable View
SPSS Processor is ready
Inicio
7. Análisis de asociaci...
Untitled - SPSS Data ...
11:03 a.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Bivariate Correlations
Variables:
habitaci
ausencia
OK
Paste
Reset
Cancel
Help
Correlation Coefficients
Pearson
Kendall's tau-b
Spearman
Test of Significance
Two-tailed
One-tailed
Flag significant correlations
Options...
Data View
Variable View
SPSS Processor is ready
Inicio
7. Análisis de asociaci...
Untitled - SPSS Data ...
11:03 a.m.


Output4 - SPSS Viewer
File Edit View Insert Format Analyze Graphs Utilities Window Help
Nonparametric Correlations
Correlations
HABITACI
AUSENCIA
Spearman's rho
HABITACI
Correlation Coefficient
1.000
.806**
Sig. (2-tailed)
.
.005
N
10
10
AUSENCIA
Correlation Coefficient
.806**
1.000
Sig. (2-tailed)
.005
.
N
10
10
**. Correlation is significant at the 0.01 level (2-tailed).
SPSS Processor is ready
Inicio
7. Análisis de asociaci...
Untitled - SPSS Data ...
Output4 - SPSS Viewer
11:04 a.m.

Según se observa en la última imagen, el valor del coeficiente de correlación por rangos de Spearman (rho) es igual a 0.806. Este valor elevado y positivo, indica que existe una fuerte relación directamente proporcional entre ambas variables, de manera que se cumple la sospecha de los miembros del departamento, pues mientras más habitaciones le asignan a las camareras por encima de su norma diaria, más ausencias presentan ellas a su puesto de trabajo (ya sean justificadas mediante certificado médico, o injustificadas), y viceversa, con un nivel de significación del 1%. Como consecuencia, cabe esperar afectaciones en el servicio de alojamiento, las cuales repercuten en mayor número de quejas de clientes externos o insatisfacción de los mismos.

7.3. Coeficiente de correlación de Pearson.

Este coeficiente ofrece, en su interpretación, dos elementos importantes:

- la intensidad de la relación entre las dos variables
- el sentido de la relación entre las dos variables

Como este coeficiente toma valores entre -1 y 1, véase las clasificaciones del sentido e intensidad de la relación:

- correlación negativa perfecta: igual a -1
- correlación negativa muy fuerte: igual a -0.90
- correlación negativa considerable: igual a -0.75
- correlación negativa media: igual a -0.50
- correlación negativa débil: igual a -0.10
- no existe correlación alguna entre las variables: igual a 0
- correlación positiva débil: igual a 0.10
- correlación positiva media: igual a 0.50
- correlación positiva considerable: igual a 0.75
- correlación positiva muy fuerte: igual a 0.90
- correlación positiva perfecta: igual a 1

Obsérvese que el número del coeficiente, indica la intensidad de la relación

entre las dos variables cuantitativas (de débil a perfecta correlación). El signo refleja el sentido (relación directamente o inversamente proporcional entre las dos variables). Véase un ejemplo.

Ejemplo 2:

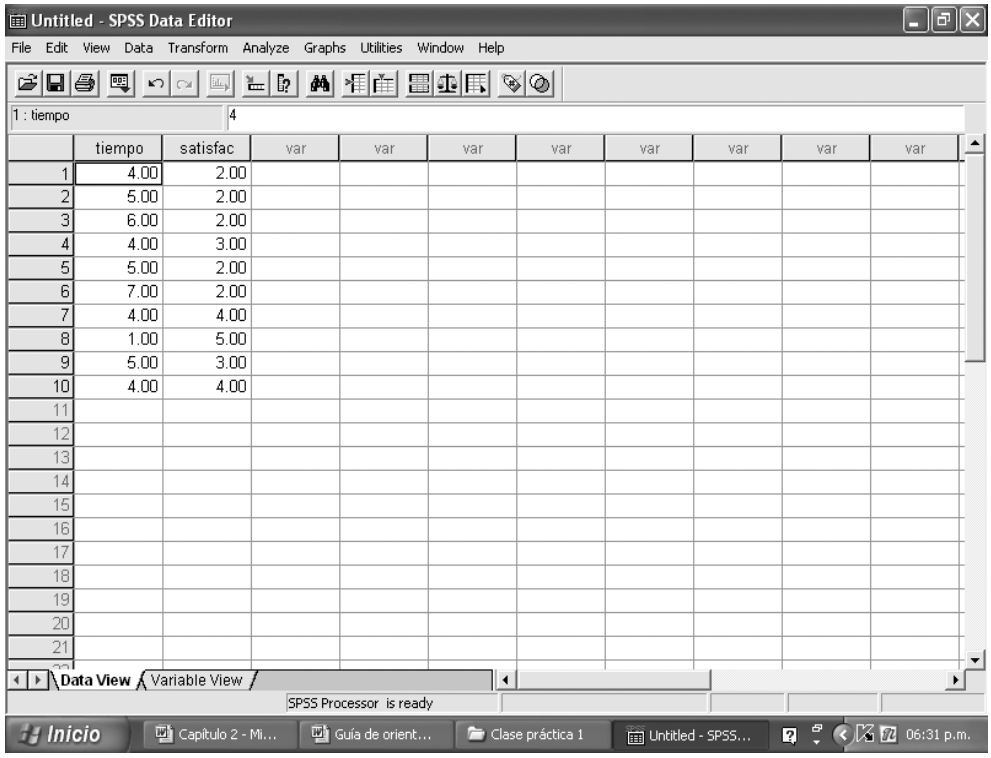
Un grupo de estudiantes de primer año de Licenciatura en Turismo, ha decidido centrar su estudio en la aparente relación que existe entre el tiempo de demora de un servicio, y el nivel de satisfacción de los clientes que consumen dicho servicio. Para ello, observaron durante un mes los tiempos de demora de cientos de clientes que transitaron por los 10 sub-procesos del servicio de Recepción en el Hotel Z, y luego aplicaron una encuesta a tales clientes para medir su nivel de satisfacción con el servicio recibido (este nivel fue medido con una escala de tipo Likert). Los datos se muestran a continuación:

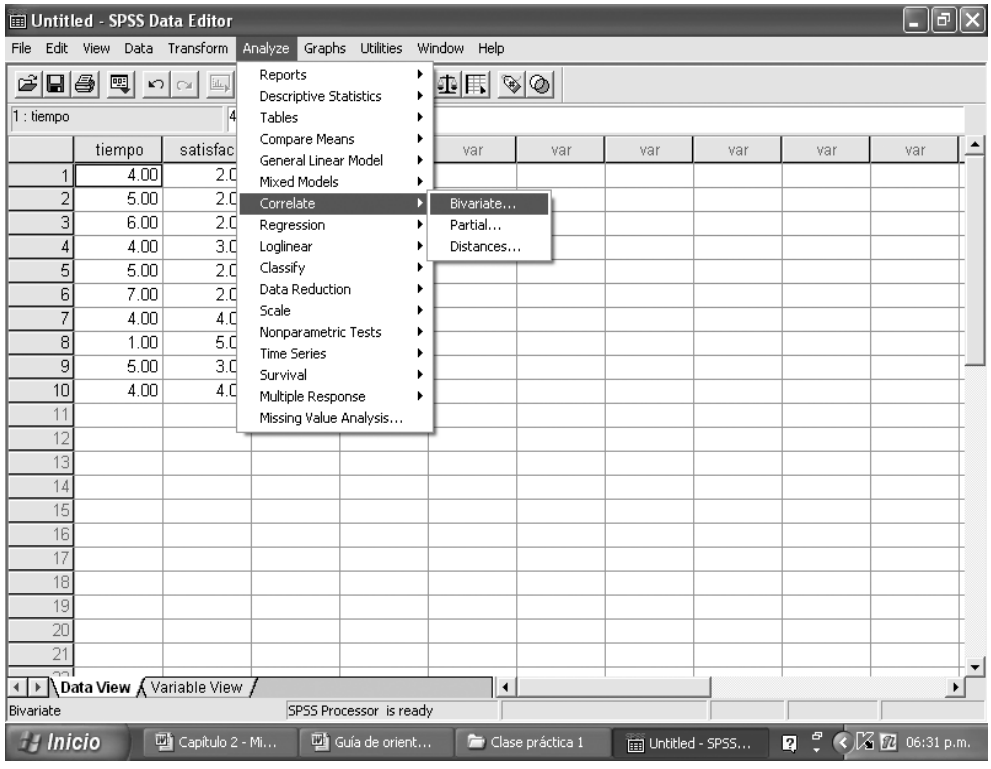
Sub-procesos	T (tiempo en min.)	S (satisfacción del cliente)
P_1	4	2
P_2	5	2
P_3	6	2
P_4	4	3
P_5	5	2
P_6	7	2
P_7	4	4
P_8	1	5
P_9	5	3
P_{10}	4	4

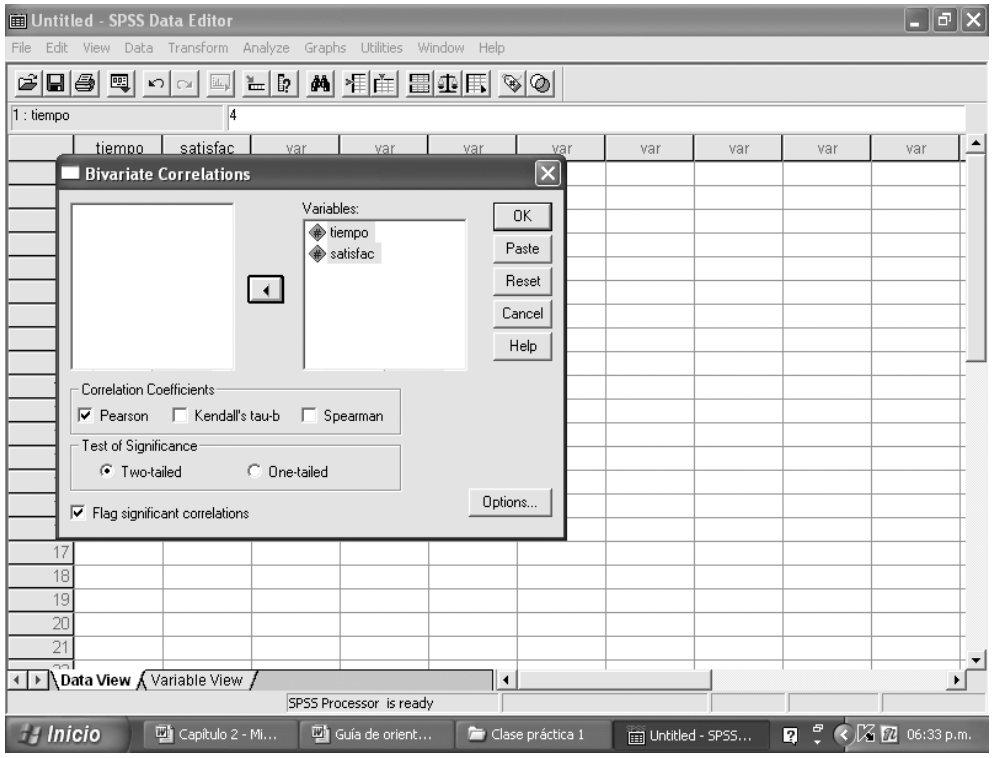
La hipótesis de los estudiantes consiste en enunciar que: “a menor tiempo de demora de cada sub-proceso, mayor es la satisfacción de los clientes” con un nivel de confiabilidad del 99%. ¿Será cierta esta sospecha?

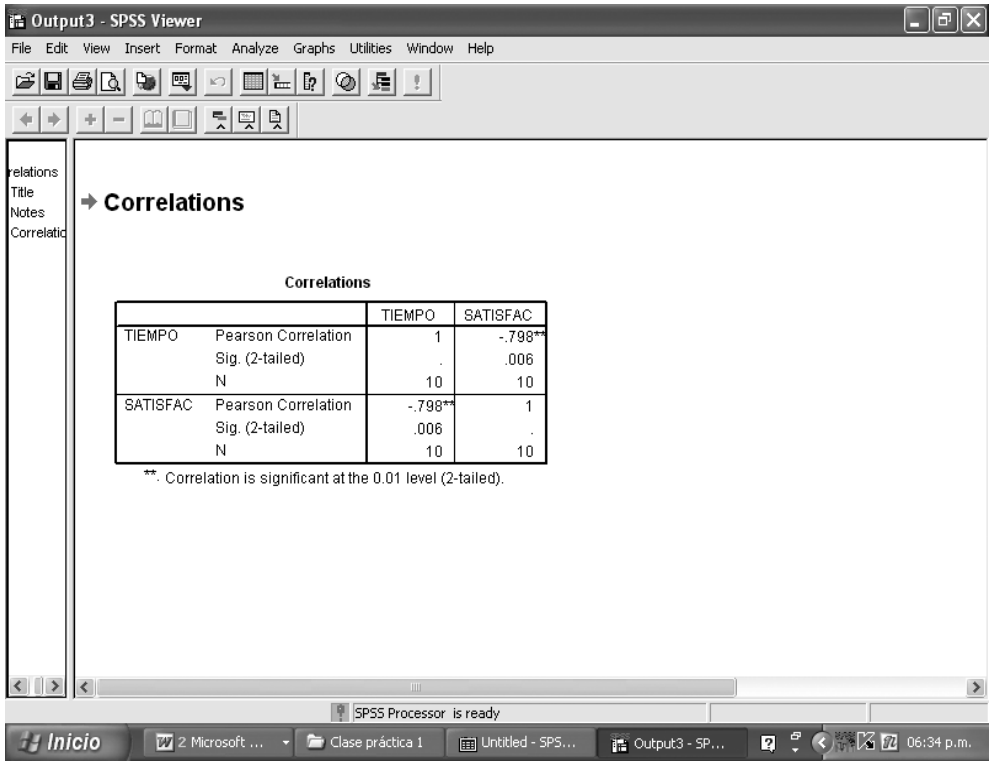
Solución:

Utilizando el SPSS, sería:


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : tiempo
4
tiempo satisfac var var var var var var var var
1 4.00 2.00
2 5.00 2.00
3 6.00 2.00
4 4.00 3.00
5 5.00 2.00
6 7.00 2.00
7 4.00 4.00
8 1.00 5.00
9 5.00 3.00
10 4.00 4.00
Data View Variable View
SPSS Processor is ready
Inicio
Capítulo 2 - Mi...
Guía de orient...
Clase práctica 1
Untitled - SPSS...
06:31 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Bivariate...
Partial...
Distances...
Data View Variable View
Bivariate
SPSS Processor is ready
Inicio
Capítulo 2 - Mi...
Guía de orient...
Clase práctica 1
Untitled - SPSS...
06:31 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : tiempo
4
tiempo satisfac var var var var var var var var
Bivariate Correlations
Variables:
tiempo
satisfac
OK
Paste
Reset
Cancel
Help
Correlation Coefficients
Pearson
Kendall's tau-b
Spearman
Test of Significance
Two-tailed
One-tailed
Flag significant correlations
Options...
17
18
19
20
21
Data View
Variable View
SPSS Processor is ready
Inicio
Capitulo 2 - Mi...
Guía de orient...
Clase práctica 1
Untitled - SPSS...
06:33 p.m.


Output3 - SPSS Viewer
File Edit View Insert Format Analyze Graphs Utilities Window Help
relations
Title
Notes
Correlatic
Correlations
Correlations
TIEMPO SATISFAC
TIEMPO Pearson Correlation 1 -.798**
Sig. (2-tailed) . .006
N 10 10
SATISFAC Pearson Correlation -.798** 1
Sig. (2-tailed) .006 .
N 10 10
**. Correlation is significant at the 0.01 level (2-tailed).
SPSS Processor is ready
Inicio
2 Microsoft ...
Clase práctica 1
Untitled - SPS...
Output3 - SP...
06:34 p.m.

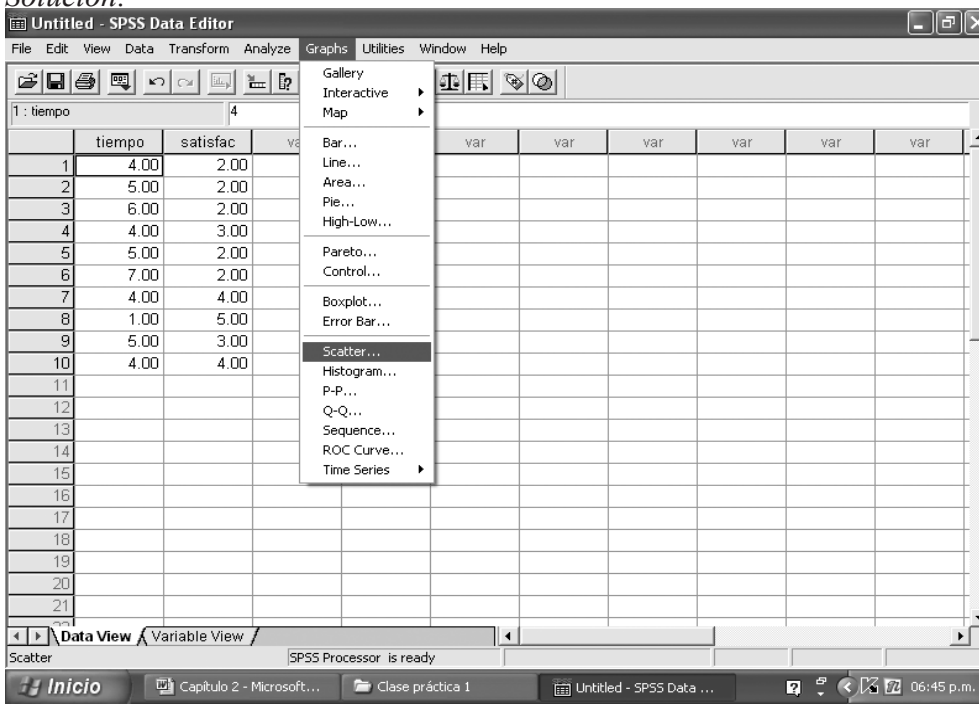
Según se observa, el coeficiente de correlación de Pearson posee un valor igual a -0.798. En primera instancia, como ese valor es diferente de cero, los estudiantes pueden afirmar que sí existe relación entre el tiempo de demora del servicio y la satisfacción del cliente. En segundo lugar, dígase que existe una relación de considerable a muy fuerte entre ambas variables. Finalmente, dicha relación es inversamente proporcional (negativa). Es así como puede resumirse que la hipótesis de los estudiantes se corrobora, pues a menor tiempo de demora de cada sub-proceso de Recepción, mayor es la satisfacción de los clientes (y viceversa) con un nivel de significación del 1%.

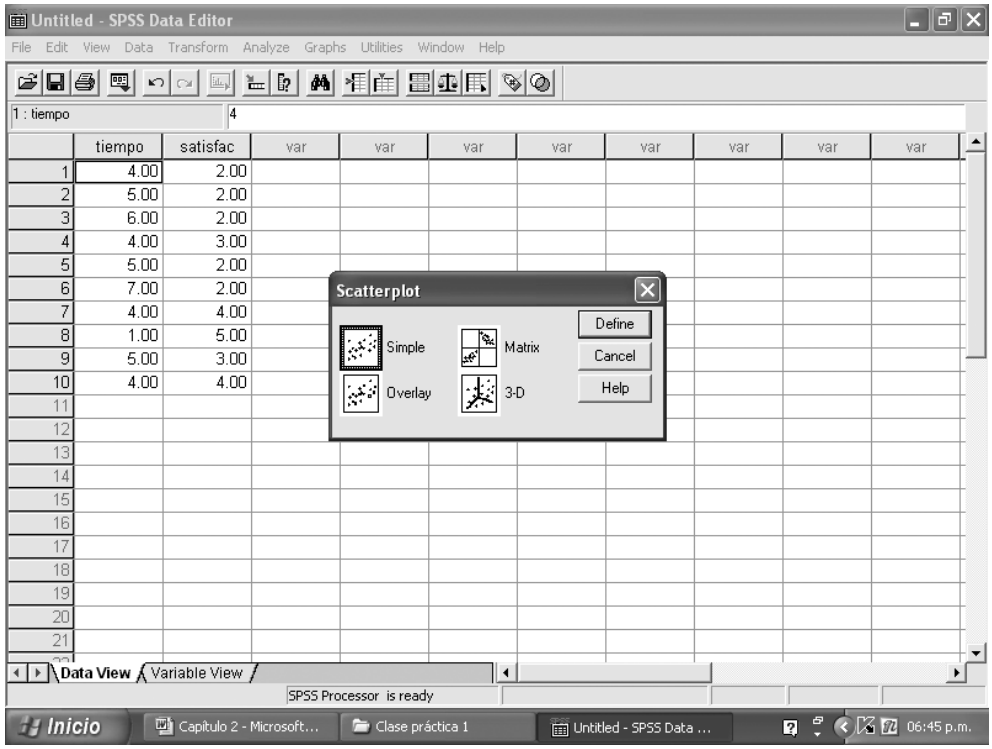
7.4. Gráfico.

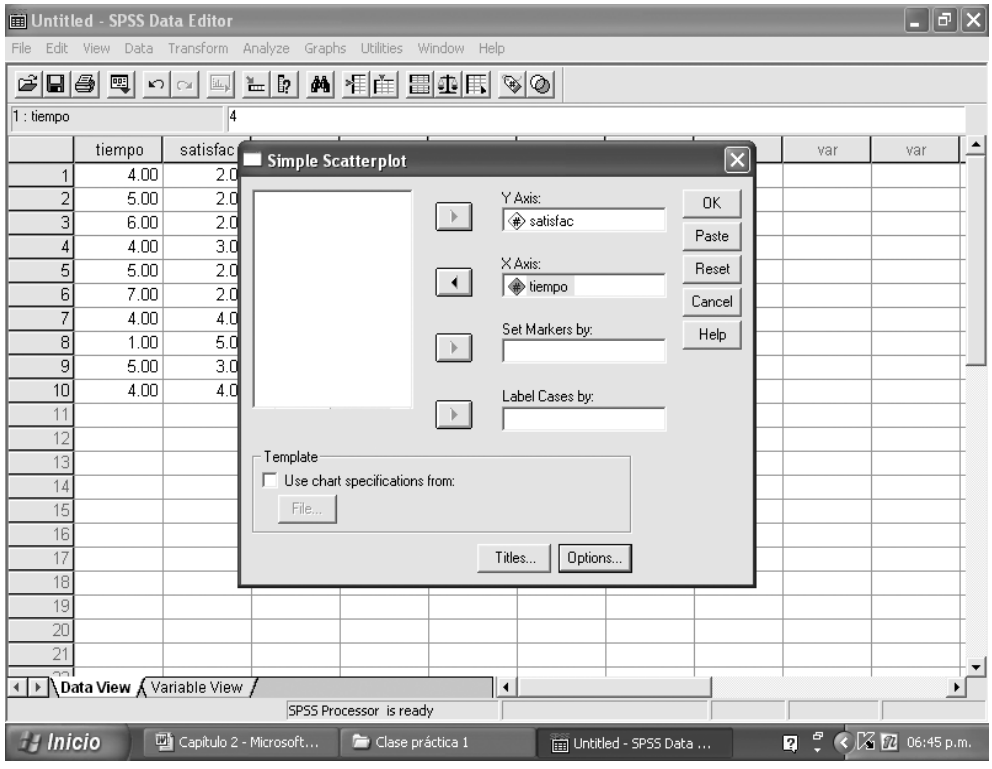
El gráfico más utilizado en las distribuciones bidimensionales, es el de dispersión.
Continuación del *ejemplo 2*:

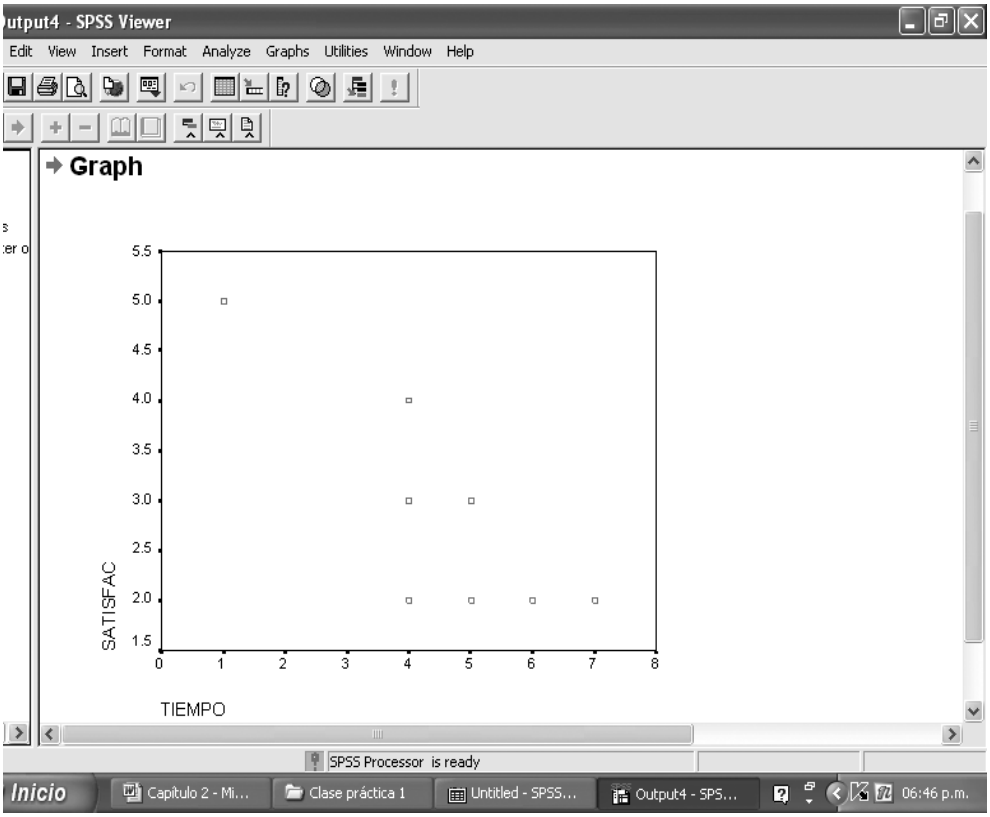
Después de demostrar que sí existe relación entre ambas variables, el grupo de estudiantes desea obtener una imagen gráfica del comportamiento de la misma en un eje de coordenadas. Para ello, es preciso elaborar un gráfico de dispersión.

Solución:




Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : tiempo
4
tiempo satisfac var var var var var var var var
1 4.00 2.00
2 5.00 2.00
3 6.00 2.00
4 4.00 3.00
5 5.00 2.00
6 7.00 2.00
7 4.00 4.00
8 1.00 5.00
9 5.00 3.00
10 4.00 4.00
Scatterplot
Simple
Matrix
Overlay
3-D
Define
Cancel
Help
Data View
Variable View
SPSS Processor is ready
Inicio
Capítulo 2 - Microsoft...
Clase práctica 1
Untitled - SPSS Data ...
06:45 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : tiempo
4
tiempo satisfac var var
Simple Scatterplot
Y Axis:
satisfac
X Axis:
tiempo
Set Markers by:
Label Cases by:
OK
Paste
Reset
Cancel
Help
Template
Use chart specifications from:
File...
Titles...
Options...
Data View
Variable View
SPSS Processor is ready
Inicio
Capítulo 2 - Microsoft...
Clase práctica 1
Untitled - SPSS Data ...
06:45 p.m.



Según la información que ofrece el gráfico, los pares ordenados de ambas variables, muestran una tendencia lineal decreciente entre las mismas (es lógico, pues el valor del coeficiente es negativo). Esto permitirá obtener una función lineal que sirva para estimar el comportamiento de la satisfacción de los clientes, sobre la base de la variabilidad que presente el tiempo de demora de los sub-procesos de Recepción.

7.5. Generalidades acerca del análisis de regresión lineal.

En Estadística, este análisis constituye uno de los más importantes y utilizados en cualquier campo de la práctica pues constantemente, los directivos, investigadores, etc., necesitan estimar el comportamiento de una variable sobre la base del estudio de la influencia que sobre ella, tiene otra o más variables diferentes. De este modo, se puede suponer un modelo o función

matemática que muestre cuánto depende el comportamiento de una variable (Y ó endógena o dependiente) en base al comportamiento de otra o más variables que influyen sobre ella (X_1, X_2…X_n ó exógenas o independientes o predictoras, como desee llamárseles).

Los modelos pueden ser logarítmicos, polinomiales, exponenciales, lineales, etc. Aquí, los ejemplos se centrarán en el modelo de regresión lineal tanto simple (una sola variable independiente) como múltiple (dos o más variables independientes) por su gran utilización en la práctica.

En el análisis de regresión, se obtiene información muy variada que facilita la toma de decisiones. El coeficiente de regresión lineal o de determinación R^2, es un indicador que aporta información valiosa. El mismo oscila entre 0 y 1 mostrando la magnitud de la influencia que sobre la variable endógena, tiene la exógena (s).

7.6. Análisis de regresión lineal simple.

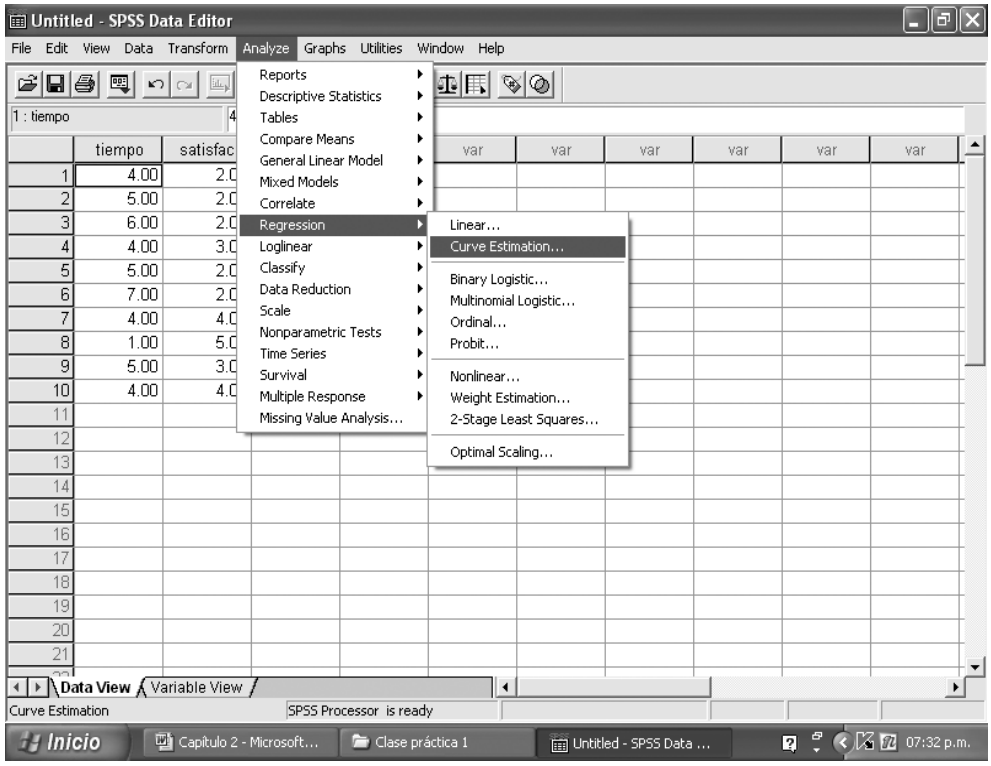
Continuando con el *ejemplo 2:*

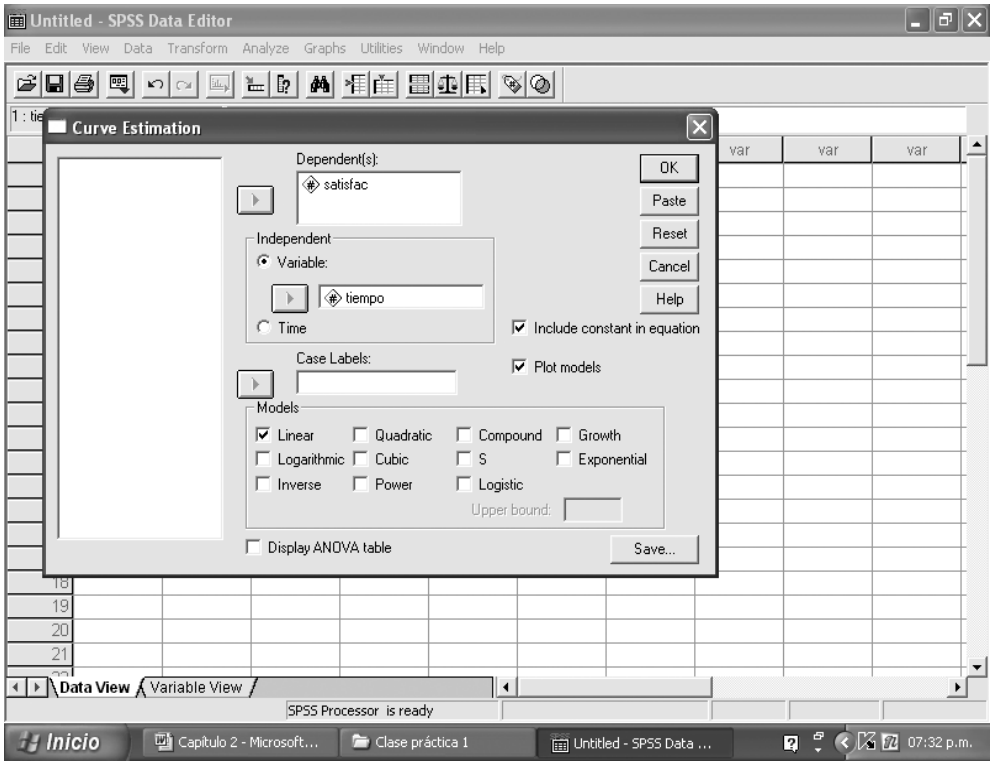
Como los estudiantes han podido comprobar que existe una relación significativa entre ambas variables, ahora se preguntan en qué medida una variación del tiempo de demora de los sub-procesos del servicio de Recepción, afecta la satisfacción de los clientes con un nivel de confiabilidad del 99%.

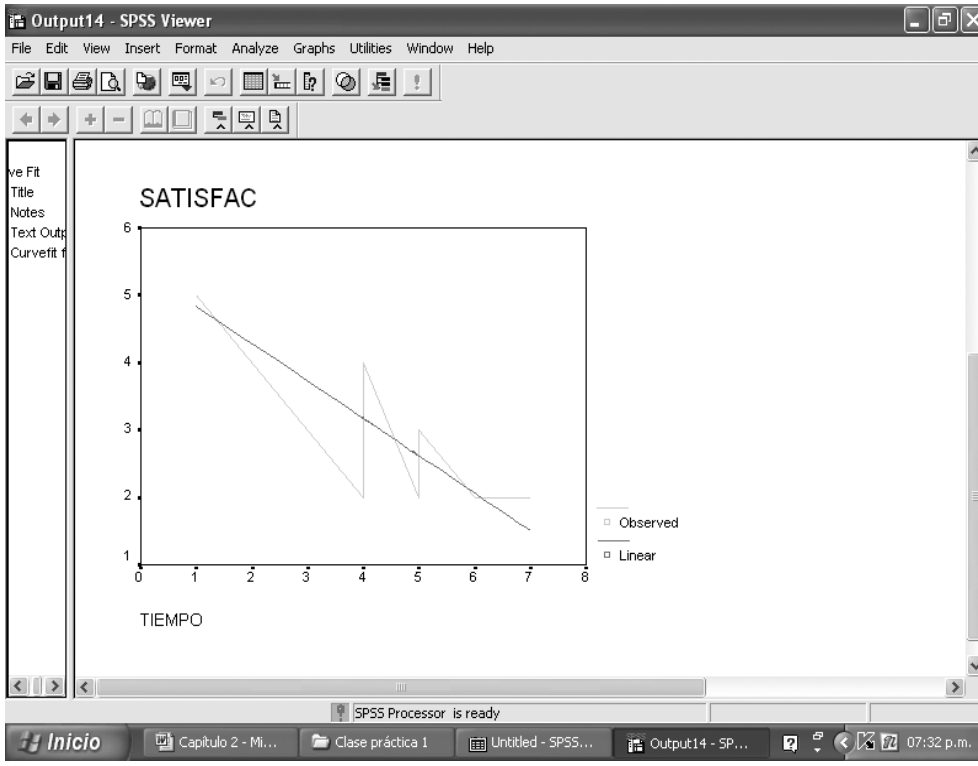
Solución:

Se buscará primero una función lineal gráfica que demuestre si el tiempo de demora del servicio, influye en la satisfacción de los clientes. En caso de existir, se procederá a representarla numéricamente.

Nota: este *ejemplo 2* requiere realizar un análisis de regresión lineal del tipo simple, pues se está estimando el comportamiento de una variable, sobre la base del estudio de la influencia que sobre ella, tiene otra (una) variable diferente.

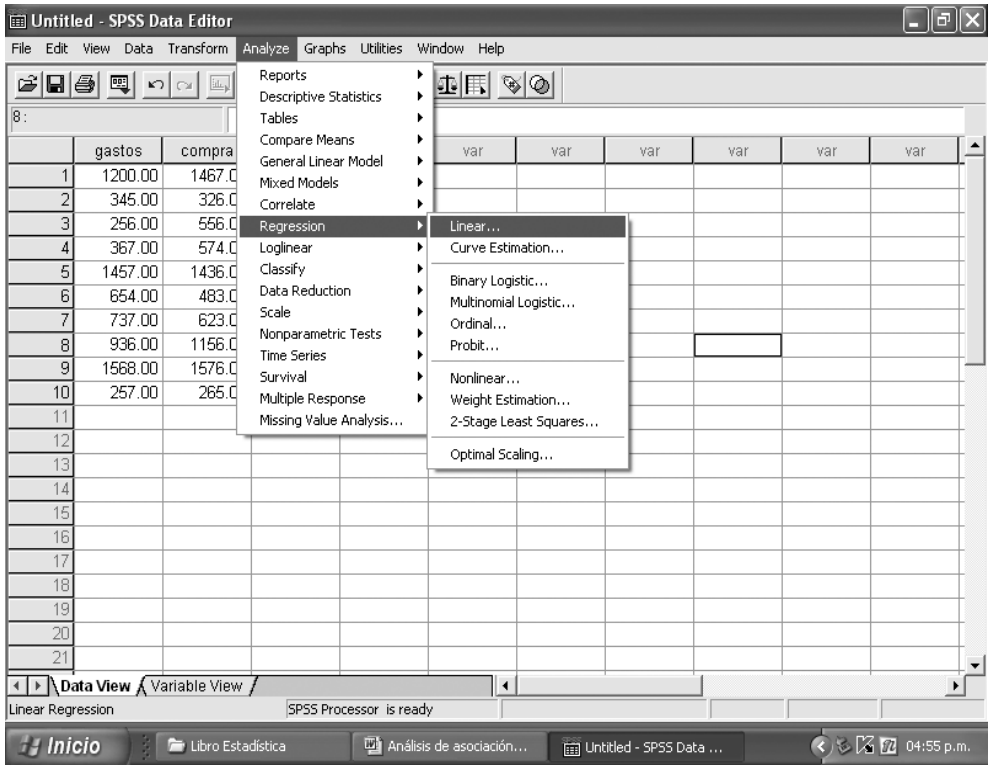

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Linear...
Curve Estimation...
Binary Logistic...
Multinomial Logistic...
Ordinal...
Probit...
Nonlinear...
Weight Estimation...
2-Stage Least Squares...
Optimal Scaling...
1 : tiempo
tiempo
satisfac
Data View
Variable View
Curve Estimation
SPSS Processor is ready
Inicio
Capítulo 2 - Microsoft...
Clase práctica 1
Untitled - SPSS Data ...
07:32 p.m.

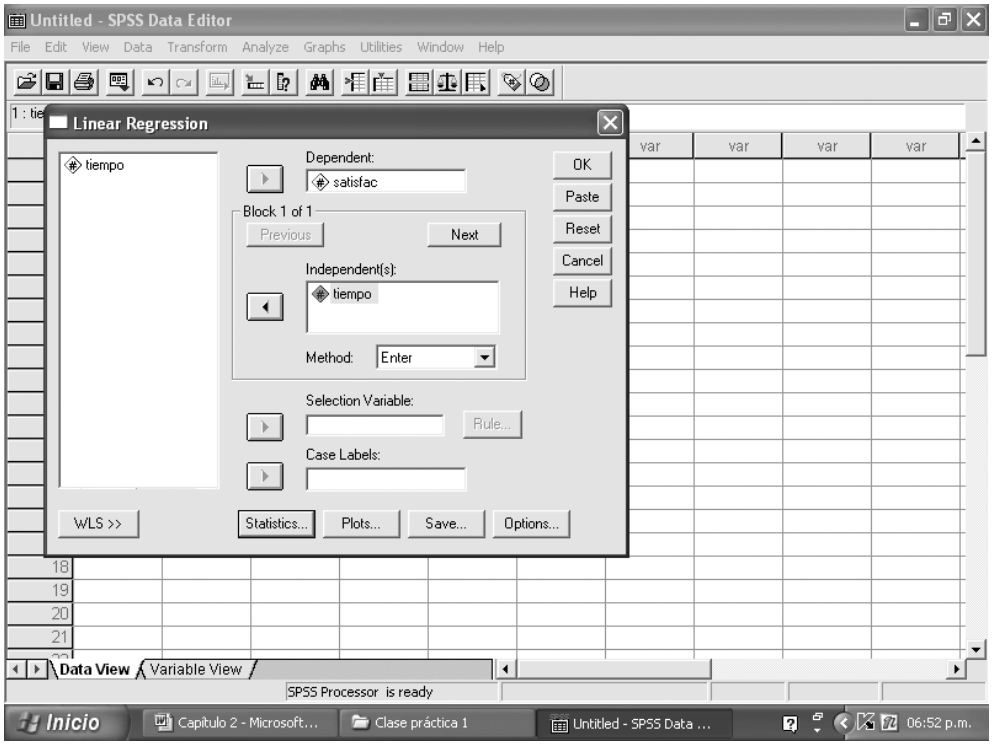

Untitled - SPSS Data Editor
Curve Estimation
Dependent(s):
satisfac
Independent
Variable:
tiempo
Time
Case Labels:
Include constant in equation
Plot models
Models
Linear
Quadratic
Compound
Growth
Logarithmic
Cubic
S
Exponential
Inverse
Power
Logistic
Upper bound:
Display ANOVA table
OK
Paste
Reset
Cancel
Help
Save...
Data View
Variable View
SPSS Processor is ready
Inicio
Capítulo 2 - Microsoft...
Clase práctica 1
Untitled - SPSS Data ...
07:32 p.m.

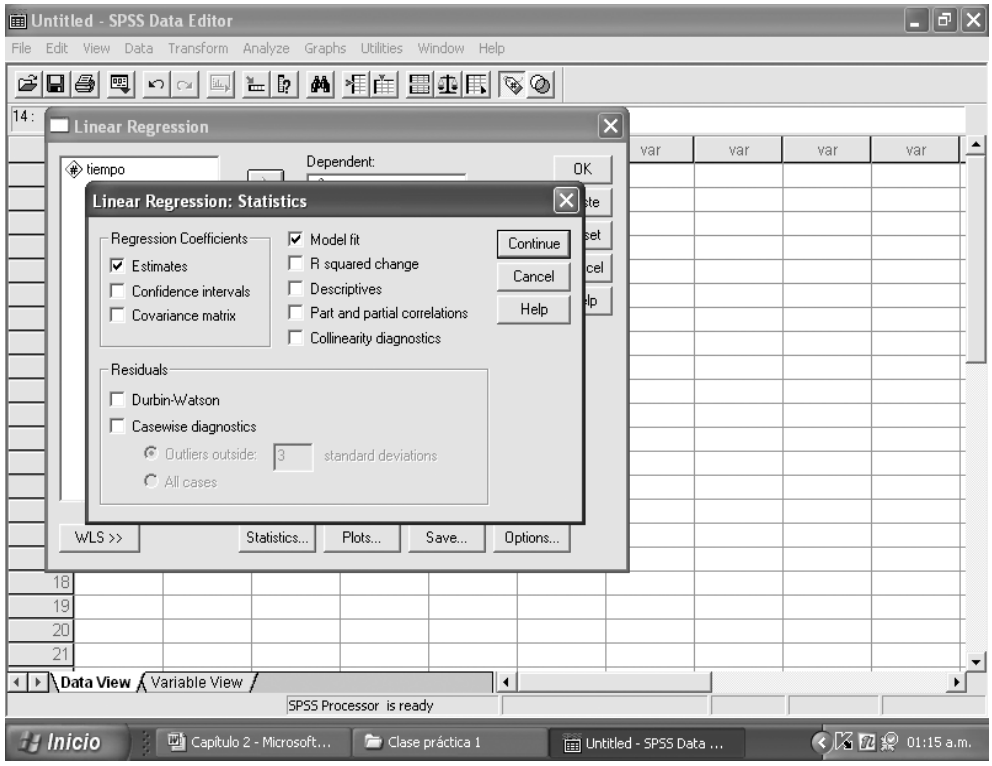


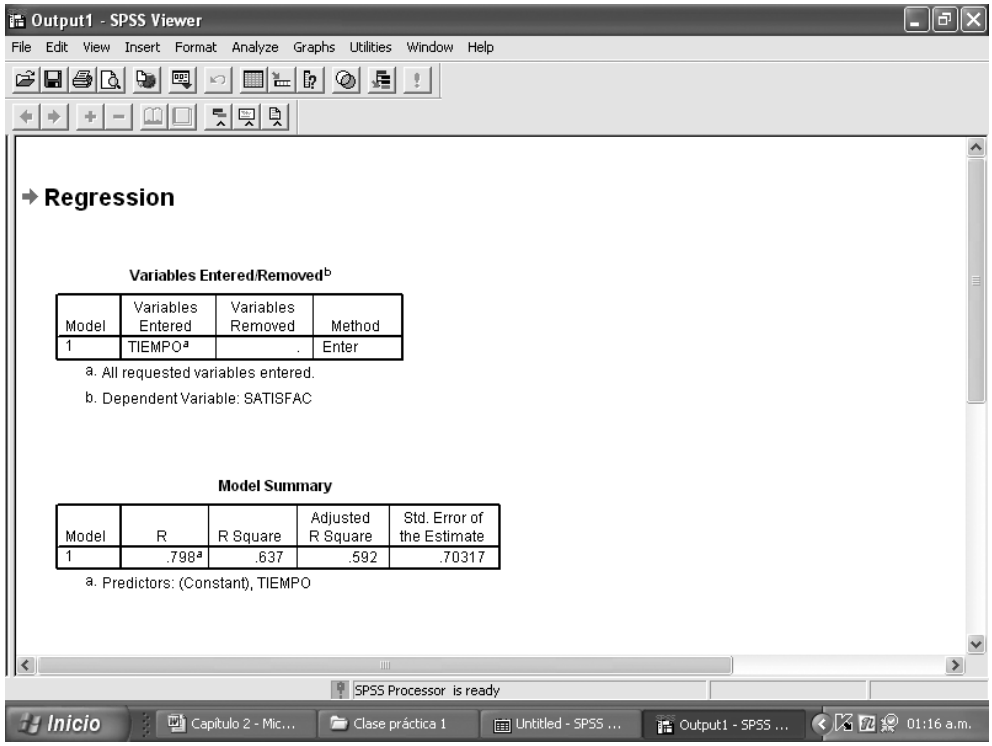
Véase que las líneas verdes son segmentos que unen los pares ordenados de ambas variables que se observan en el gráfico de dispersión, y la línea roja es la tendencia promedio de esos pares ordenados o función lineal. Que la función lineal sea monótona decreciente, corrobora que a menor tiempo de demora del servicio, mayor es la satisfacción de los clientes.

Prosiguiendo:

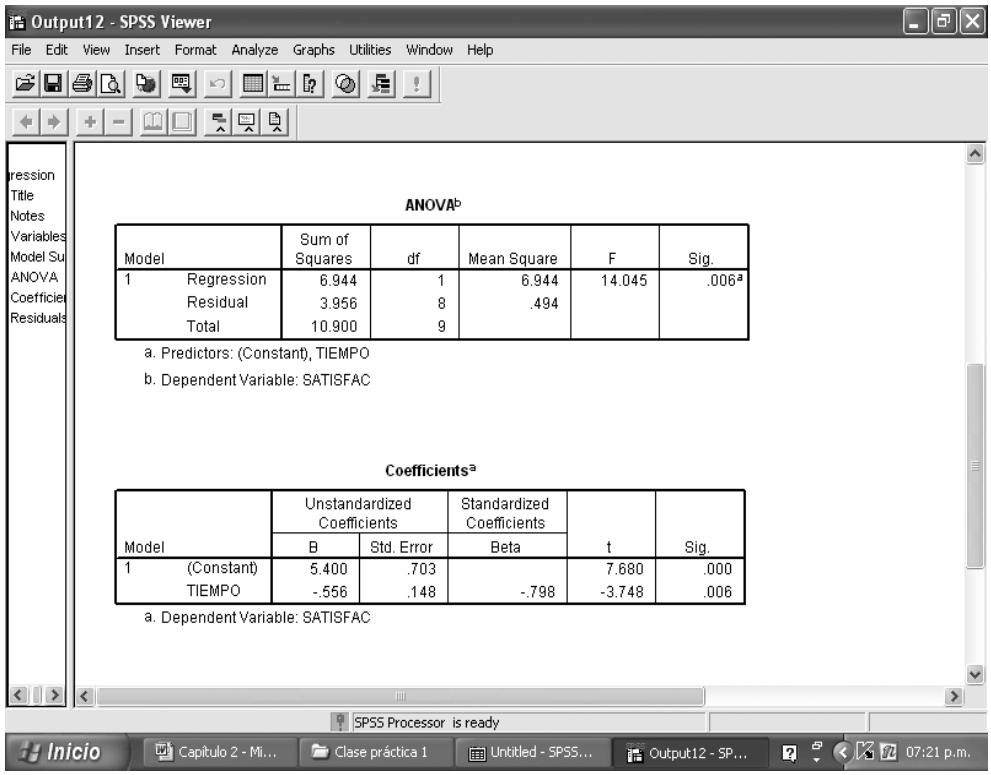

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Linear...
Curve Estimation...
Binary Logistic...
Multinomial Logistic...
Ordinal...
Probit...
Nonlinear...
Weight Estimation...
2-Stage Least Squares...
Optimal Scaling...
gastos
compra
Data View
Variable View
Linear Regression
SPSS Processor is ready
Inicio
Libro Estadística
Análisis de asociación...
Untitled - SPSS Data ...
04:55 p.m.


Untitled - SPSS Data Editor
Linear Regression
tiempo
Dependent:
satisfac
Block 1 of 1
Previous
Next
Independent(s):
tiempo
Method:
Enter
Selection Variable:
Rule...
Case Labels:
WLS >>
Statistics...
Plots...
Save...
Options...
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready
Inicio
Capítulo 2 - Microsoft...
Clase práctica 1
Untitled - SPSS Data ...
06:52 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Linear Regression
tiempo
Dependent:
OK
Linear Regression: Statistics
Regression Coefficients
Estimates
Confidence intervals
Covariance matrix
Model fit
R squared change
Descriptives
Part and partial correlations
Collinearity diagnostics
Continue
Cancel
Help
Residuals
Durbin-Watson
Casewise diagnostics
Outliers outside:
3
standard deviations
All cases
WLS >>
Statistics...
Plots...
Save...
Options...
var
Data View
Variable View
SPSS Processor is ready
Inicio
Capítulo 2 - Microsoft...
Clase práctica 1
Untitled - SPSS Data ...
01:15 a.m.


Output1 - SPSS Viewer
File Edit View Insert Format Analyze Graphs Utilities Window Help
Regression
Variables Entered/Removed[b]
Model | Variables Entered | Variables Removed | Method
1 | TIEMPO[a] | . | Enter
a. All requested variables entered.
b. Dependent Variable: SATISFAC
Model Summary
Model | R | R Square | Adjusted R Square | Std. Error of the Estimate
1 | .798[a] | .637 | .592 | .70317
a. Predictors: (Constant), TIEMPO
SPSS Processor is ready
Inicio
Capítulo 2 - Mic...
Clase práctica 1
Untitled - SPSS ...
Output1 - SPSS ...
01:16 a.m.

El software muestra una salida amplia. En la primera parte se observa que el valor del coeficiente de determinación R^2 es igual a 0.637. Esto indica que una variabilidad del tiempo de demora del servicio (variable independiente, exógena o predictora) influye en un 64% en la satisfacción de los clientes (variable dependiente o endógena).



La tabla ANOVA hace referencia a la dócima de la pendiente:

H_0: $B_1 = 0$ (pendiente igual a cero)

H_1: $B_1 \neq 0$ (pendiente diferente de cero)

Como el valor de probabilidad de esta dócima es igual a 0.006 y dicho valor es menor que 0.01, entonces se cumple la región crítica y se rechaza la hipótesis nula, lo cual indica que la pendiente es diferente de cero y que la satisfacción de los clientes sí depende del tiempo de demora del servicio. Más abajo se observa que, efectivamente, la pendiente (b_1) toma un valor igual a -0.556 (negativo al igual que el coeficiente de correlación de Pearson anteriormente) y el intercepto (b_0) igual a 5.4.

Con estos valores se obtiene la ecuación de regresión lineal simple que sirve para hacer estimaciones del comportamiento de la variable dependiente Y (satisfacción del cliente) según la variabilidad que presente la variable independiente X (tiempo de demora del servicio).

La ecuación sería:

$Y = -0.556 X + 5.4$

7.7. Análisis de regresión lineal múltiple.

Véase un ejemplo.

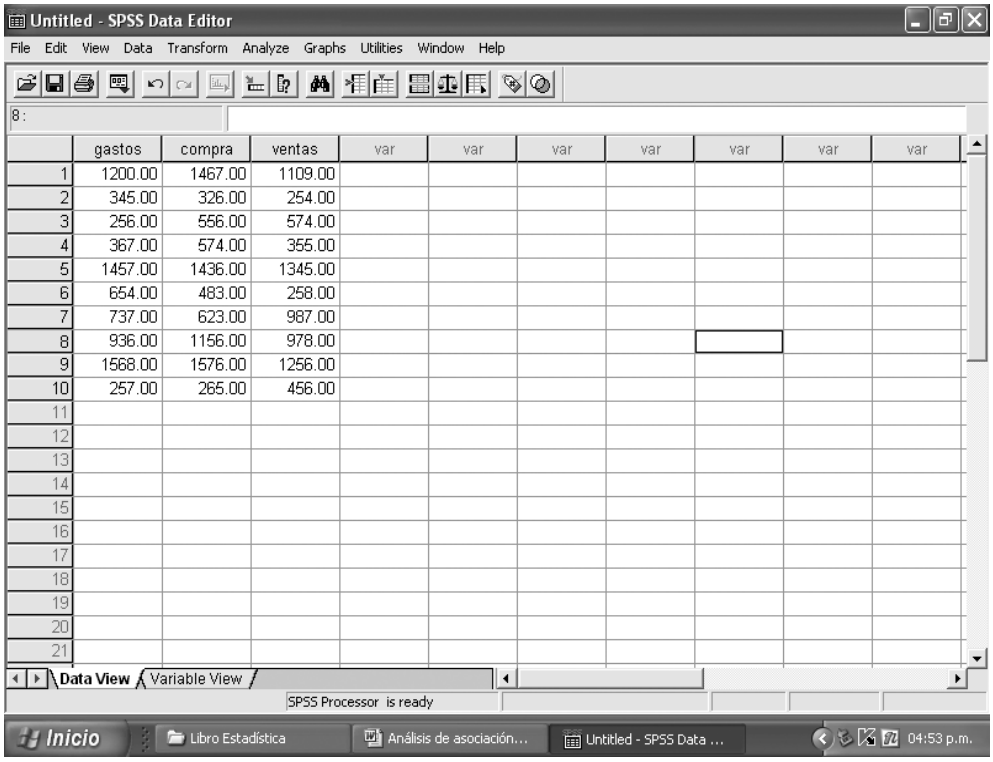
Ejemplo 3:

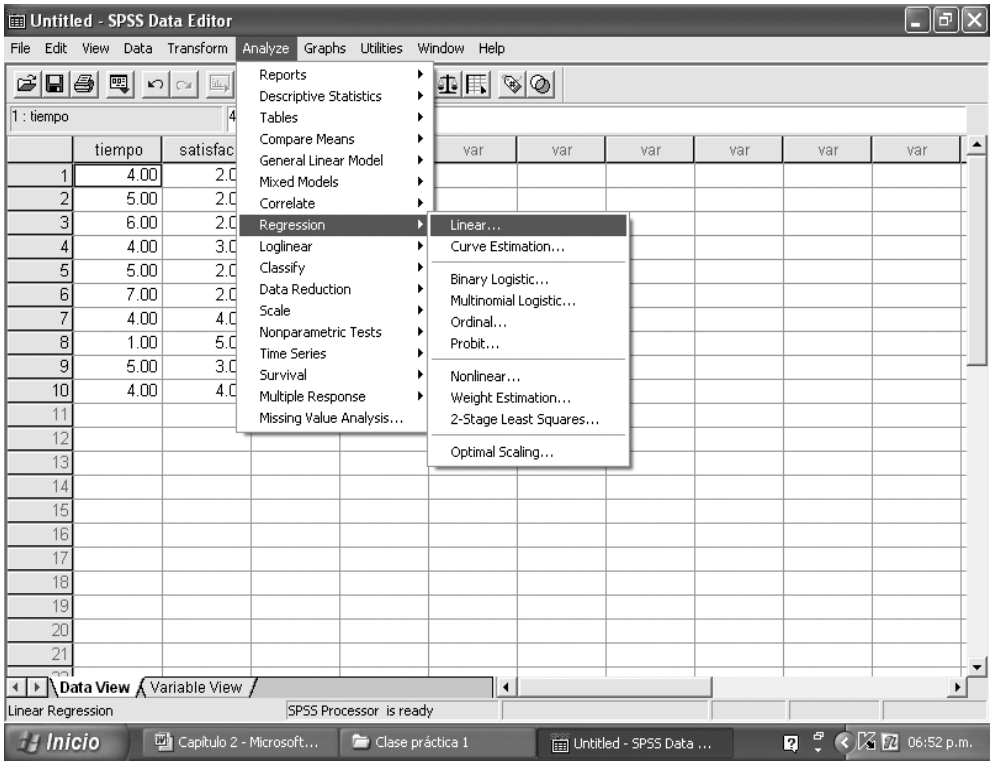
La Agencia de Viajes N desea conocer el impacto que tienen los gastos de promoción y la compra de medios materiales, en niveles futuros de venta de dicha empresa. Para ello, la dirección decidió tomar una muestra de 10 días analizando diariamente los niveles de gastos, los de compra y realizando un pronóstico de las ventas futuras con un nivel de significación del 5%.

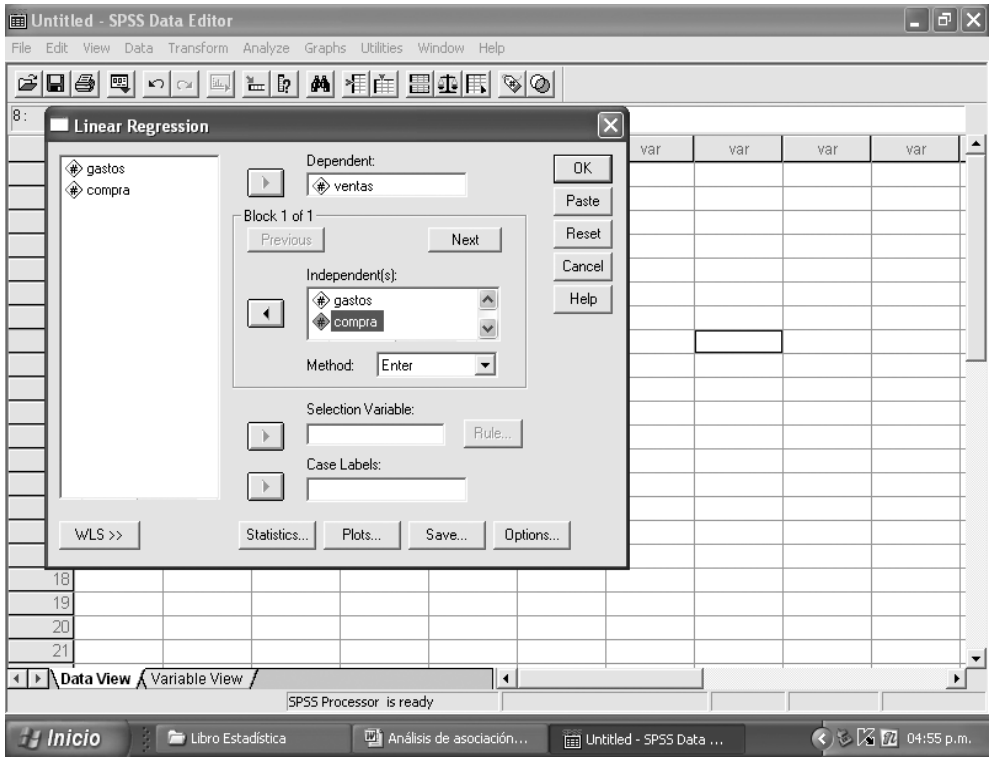
Los datos se observan en la tabla que aparece a continuación:

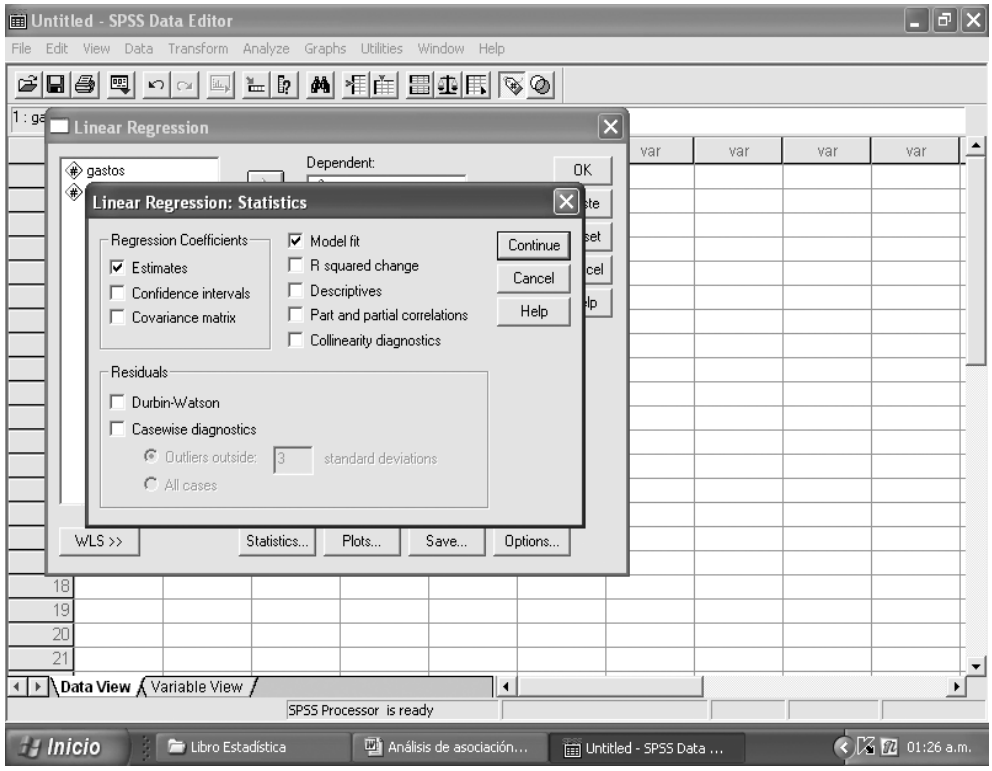
Gastos de promoción ($)	Compra de medios materiales ($)	Ventas pronosticadas ($)
1200.00	1467.00	1109.00
345.00	326.00	254.00
256.00	556.00	574.00
367.00	574.00	355.00
1457.00	1436.00	1345.00
654.00	483.00	258.00
737.00	623.00	987.00
936.00	1156.00	978.00
1568.00	1576.00	1256.00
257.00	265.00	456.00

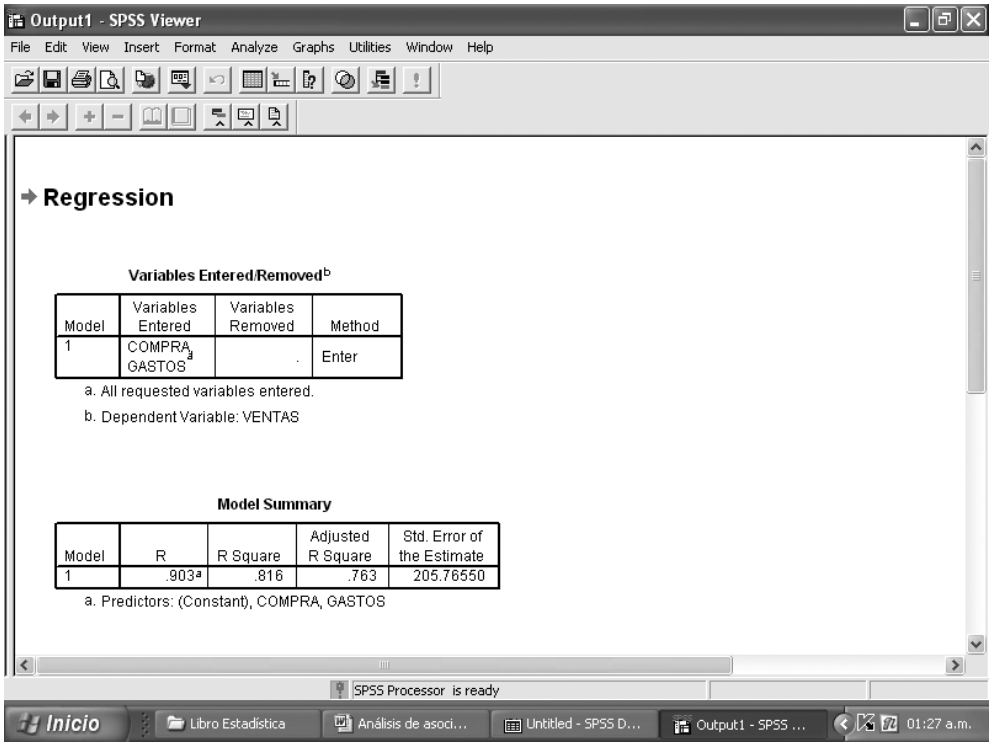
Solución:

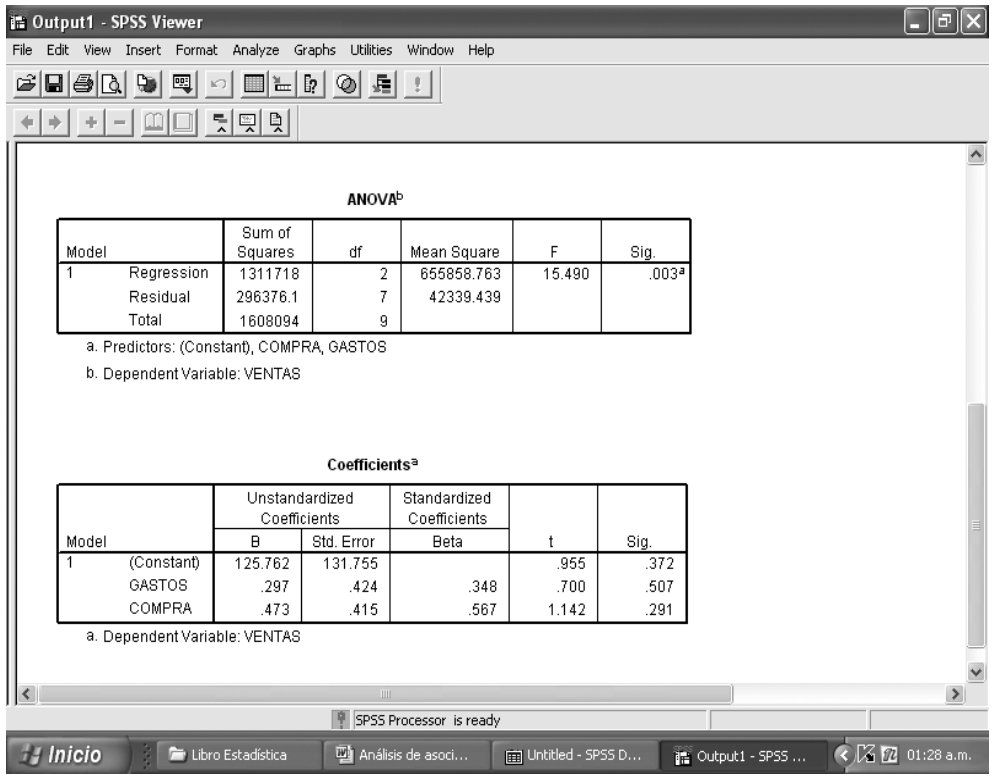

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
8 :
gastos compra ventas var var var var var var var
1 1200.00 1467.00 1109.00
2 345.00 326.00 254.00
3 256.00 556.00 574.00
4 367.00 574.00 355.00
5 1457.00 1436.00 1345.00
6 654.00 483.00 258.00
7 737.00 623.00 987.00
8 936.00 1156.00 978.00
9 1568.00 1576.00 1256.00
10 257.00 265.00 456.00
Data View Variable View
SPSS Processor is ready
Inicio
Libro Estadística
Análisis de asociación...
Untitled - SPSS Data ...
04:53 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Linear...
Curve Estimation...
Binary Logistic...
Multinomial Logistic...
Ordinal...
Probit...
Nonlinear...
Weight Estimation...
2-Stage Least Squares...
Optimal Scaling...
1 : tiempo
tiempo satisfac var var var var var var
1 4.00 2.0
2 5.00 2.0
3 6.00 2.0
4 4.00 3.0
5 5.00 2.0
6 7.00 2.0
7 4.00 4.0
8 1.00 5.0
9 5.00 3.0
10 4.00 4.0
Data View Variable View
Linear Regression
SPSS Processor is ready
Inicio
Capítulo 2 - Microsoft...
Clase práctica 1
Untitled - SPSS Data ...
06:52 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Linear Regression
gastos
compra
Dependent:
ventas
Block 1 of 1
Previous
Next
Independent(s):
gastos
compra
Method:
Enter
Selection Variable:
Rule...
Case Labels:
WLS >>
Statistics...
Plots...
Save...
Options...
OK
Paste
Reset
Cancel
Help
var
Data View
Variable View
SPSS Processor is ready
Inicio
Libro Estadística
Análisis de asociación...
Untitled - SPSS Data ...
04:55 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Linear Regression
Linear Regression: Statistics
Regression Coefficients
Estimates
Confidence intervals
Covariance matrix
Model fit
R squared change
Descriptives
Part and partial correlations
Collinearity diagnostics
Continue
Cancel
Help
Residuals
Durbin-Watson
Casewise diagnostics
Outliers outside:
3
standard deviations
All cases
WLS >>
Statistics...
Plots...
Save...
Options...
Data View
Variable View
SPSS Processor is ready
Inicio
Libro Estadística
Análisis de asociación...
Untitled - SPSS Data ...
01:26 a.m.


Output1 - SPSS Viewer
File Edit View Insert Format Analyze Graphs Utilities Window Help
Regression
Variables Entered/Removed[b]
Model | Variables Entered | Variables Removed | Method
1 | COMPRA, GASTOS[a] | . | Enter
a. All requested variables entered.
b. Dependent Variable: VENTAS
Model Summary
Model | R | R Square | Adjusted R Square | Std. Error of the Estimate
1 | .903[a] | .816 | .763 | 205.76550
a. Predictors: (Constant), COMPRA, GASTOS
SPSS Processor is ready
Inicio
Libro Estadística
Análisis de asoci...
Untitled - SPSS D...
Output1 - SPSS ...
01:27 a.m.


Output1 - SPSS Viewer
File Edit View Insert Format Analyze Graphs Utilities Window Help
ANOVA[b]
Model | | Sum of Squares | df | Mean Square | F | Sig.
1 | Regression | 1311718 | 2 | 655858.763 | 15.490 | .003[a]
Residual | 296376.1 | 7 | 42339.439
Total | 1608094 | 9
a. Predictors: (Constant), COMPRA, GASTOS
b. Dependent Variable: VENTAS
Coefficients[a]
Model | | Unstandardized Coefficients B | Unstandardized Coefficients Std. Error | Standardized Coefficients Beta | t | Sig.
1 | (Constant) | 125.762 | 131.755 | | .955 | .372
GASTOS | .297 | .424 | .348 | .700 | .507
COMPRA | .473 | .415 | .567 | 1.142 | .291
a. Dependent Variable: VENTAS
SPSS Processor is ready
Inicio
Libro Estadística
Análisis de asoci...
Untitled - SPSS D...
Output1 - SPSS ...
01:28 a.m.

Obsérvese que el valor del coeficiente de determinación R^2 en la antepenúltima imagen, es igual a 0.816 indicando que el 82% de la variabilidad en las ventas, se puede predecir sobre la base de la influencia que sobre ellas tienen, los gastos de promoción y la compra de medios materiales.

La dócima de la pendiente en la penúltima imagen, refleja en la tabla ANOVA un valor de probabilidad igual a 0.003. Como dicho valor es menor que 0.05, entonces se puede afirmar que las dos variables exógenas (gastos de promoción y compra de medios materiales) son capaces de predecir el comportamiento de la variable endógena (ventas futuras), mediante un modelo de regresión lineal múltiple con un nivel de confiabilidad del 95%. La ecuación es la siguiente:

$$Y = 0.297\ X_1 + 0.473\ X_2 + 125.762$$

Los valores del intercepto b_0 (125.762), la pendiente b_1 que acompaña a la primera variable independiente (0.297) y la pendiente b_2 que acompaña a la segunda variable independiente (0.473), son igualmente tomados de los resultados que muestra el SPSS en la última imagen.

EJERCITACIÓN

En el Hotel S, un equipo de comerciales está llevando a cabo un estudio detallado del mercado canadiense, su principal emisor. En la fase inicial de análisis, desean demostrar si el nivel de ingresos percibido por los canadienses en su país, está relacionado con el nivel de gastos en el destino turístico cuando deciden viajar. Esto resulta de gran importancia para analizar cómo llevar a cabo una estrategia diferenciada y certera de venta hacia ese mercado e intentar, por tanto, elevar los ingresos del hotel y el destino turístico en general. Bajo un nivel de confiabilidad del 99%, los datos son los siguientes:

	Ingresos en el país ($)	Gastos en el destino ($)
1	7868.90	2657.45
2	6849.80	1647.60
3	7584.69	2758.35
4	8768.80	3856.45
5	6879.70	1758.30
6	5342.60	1386.40
7	6748.75	1657.35
8	9577.85	3869.45
9	7685.55	2657.60
10	7365.00	2753.20
11	8453.70	3856.25
12	6298.00	1657.35

SOLUCIÓN

El coeficiente de correlación de Pearson es igual a 0.933 por lo que en primera instancia, el equipo de comerciales puede afirmar que sí existe una relación muy fuerte y directamente proporcional entre los ingresos de los canadienses en su país, y los gastos en el destino turístico con un nivel de significación del 1%.

En segunda instancia, el equipo ha detectado que el coeficiente de determinación lineal, es igual a 0.87, indicando que una variabilidad en los niveles de ingresos de los canadienses en su país, repercute en un 87% en los niveles de gastos de los mismos en el destino turístico.

Como el valor de probabilidad de la dócima de la pendiente es igual a 0.000, entonces el equipo puede afirmar que sí existe grado de dependencia de los gastos sobre los ingresos con un nivel de seguridad del 99%, tal y como muestran los resultados del coeficiente de determinación lineal.

Por último, la ecuación que servirá a los comerciales para estimar los niveles futuros de gastos de los canadienses cuando arriben al hotel o al destino turístico, siguiendo un modelo de regresión lineal, es la siguiente:

$Y = 0.76\ X - 3131.14$

Capítulo
8

Análisis factorial.

8.1. Concepto de análisis factorial.

El análisis factorial es una técnica de reducción de datos que sirve para encontrar grupos homogéneos de variables, a partir de un conjunto numeroso de variables. Esos grupos homogéneos se forman con las variables que correlacionan mucho entre sí y procurando, inicialmente, que unos grupos sean independientes de otros.

Como el análisis factorial es una técnica de reducción de la dimensionalidad de los datos, su propósito último consiste en buscar el número mínimo de dimensiones capaces de explicar el máximo de información contenida en los datos.

Debe señalarse que todas las variables que intervienen en el análisis, cumplen el mismo papel, o sea, todas son independientes en el sentido de que no existe a priori, una dependencia conceptual de unas con otras.

Por último, vale destacar que la importancia del análisis factorial, radica también en reflejar el conjunto de variables con el menor número de factores posible, y que, a su vez, éstos tengan una interpretación clara y un sentido preciso.

8.2. Algunas puntualizaciones de interés acerca del análisis factorial.

Para que el análisis factorial conduzca a una solución eficiente, es necesario que la matriz de correlaciones entre las variables, contenga grupos de variables que correlacionen fuertemente entre sí. Esto se reflejaría en niveles

de significación pequeños muy cercanos a cero, que sería lo deseable.

En la matriz de correlaciones entre las variables, el valor del determinante de la misma, debe ser preferiblemente bien bajo pero distinto de cero. Esto corroboraría que las variables de la matriz, están linealmente relacionadas.

La matriz inversa de la matriz de correlaciones, sólo se puede obtener cuando el valor del determinante es distinto de cero. En caso de presentarse un determinante igual a cero, no será posible obtener la matriz inversa y por tanto, no se podrán emplear métodos de extracción de factores tales como: ejes principales, máxima verosimilitud, etc.
La medida de adecuación muestral de Kaiser-Meyer-Olkin (KMO) permite valorar la bondad de ajuste de los datos analizados a un modelo factorial. El estadístico KMO toma valores entre 0 y 1. Sería ideal obtener un valor de KMO elevado, aunque siempre que se halle por encima de 0.60, indica que se puede utilizar el análisis factorial con los datos que se tienen. Véase lo siguiente:

Valor de:
KMO menor que 0.50: se considera muy malo
KMO entre 0.50 y 0.59: se considera malo
KMO entre 0.60 y 0.69: se considera regular
KMO entre 0.70 y 0.79: se considera aceptable
KMO entre 0.80 y 0.89: se considera bueno
KMO entre 0.90 y 1: se considera excelente

La prueba de esfericidad de Bartlett plantea las hipótesis:

H_0: la matriz de correlaciones es una matriz de identidad
H_1: la matriz de correlaciones es una matriz de no identidad

La matriz de correlaciones se considera que es una matriz de identidad, cuando en la diagonal principal los coeficientes de correlación, son iguales a 1, y los externos a la diagonal, iguales a 0, indicando que no existe relación alguna entre las variables.

El estadígrafo de esta prueba es una transformación del determinante de la matriz de correlaciones. Lo recomendable es poder rechazar la hipótesis nula, indicando que dicha matriz, es de no identidad, lo cual quiere decir que el modelo factorial, es adecuado para explicar los datos.

La matriz de correlaciones anti-imagen, debe mostrar coeficientes de correlación pequeños entre cada par de variables, evidenciando que estas últimas comparten gran cantidad de información debido a la presencia de factores comunes, lo cual es un síntoma de idoneidad del análisis factorial.

En la tabla de comunalidades, se asume que si una variable está muy relacionada con las restantes variables del análisis, tenderá entonces a compartir su información en un factor común. La que posea un coeficiente de correlación muy bajo, debería ser excluida del análisis factorial.

La tabla de porcentajes de varianza explicada, permitirá observar cuántos factores ha extraído el programa con los datos iniciales. También el investigador puede definir cuántos factores desea extraer, colocando en el programa la cantidad deseable.

Existen varios métodos de extracción de factores, y también varios métodos de rotación de la solución factorial. Según el método de extracción seleccionado, varían los resultados del análisis, aunque no de forma notable. Igualmente sucede al escoger uno u otro método de rotación.

Cuando el análisis factorial ha sido fructífero con determinados datos iniciales, la matriz residual debe mostrar coeficientes muy pequeños donde la mayoría de los mismos, se encuentre cercano a cero.
Como la finalidad última del análisis factorial es reducir un gran número de variables a un pequeño grupo de factores, es a veces aconsejable estimar las puntuaciones factoriales para cada sujeto.

Véase un ejemplo.

Ejemplo 1:

En la Empresa de Transporte Turístico Z el Departamento de Recursos Humanos ha recopilado los datos pertenecientes a 15 trabajadores. Los mismos se distribuyen en nueve variables que son:

- nivel educativo
- categoría laboral
- salario actual
- salario inicial
- meses desde el comienzo en el puesto
- años de experiencia previa
- edad
- clasificación étnica
- sexo

Se desea entonces, bajo un nivel de confiabilidad del 95%, encontrar un número mínimo de dimensiones (factores), con el objetivo de reducir los datos iniciales partiendo de la similitud que existe entre unas y otras variables anteriores. Véase la tabulación de los datos:

	niveduca	categlab	salaract	salarini	mescopue	añexppre	edad	clasetni	sexo
1	1	1	325.00	325.00	9	10	38	1	2
2	2	3	642.00	560.00	24	12	45	3	2
3	1	1	410.00	390.00	6	8	29	2	1
4	2	2	512.00	480.00	15	7	32	1	1
5	2	3	680.00	600.00	11	9	49	2	2
6	1	1	315.00	315.00	4	3	26	1	1
7	1	2	422.00	400.00	10	2	31	3	2
8	2	3	638.00	600.00	19	7	41	3	1
9	1	1	380.00	380.00	8	6	38	2	1
10	1	2	463.00	430.00	4	8	38	1	1
11	2	2	410.00	410.00	12	5	27	2	1
12	2	2	432.00	400.00	9	5	31	2	2
13	2	3	615.00	590.00	15	12	48	1	1
14	2	2	425.00	410.00	10	6	39	1	2
15	1	1	360.00	340.00	9	8	43	3	2

Leyenda:

Nivel educativo:

1: nivel medio

2: nivel superior

Categoría laboral:

1: técnico

2: especialista

3: directivo

Clasificación étnica:

1: blanco

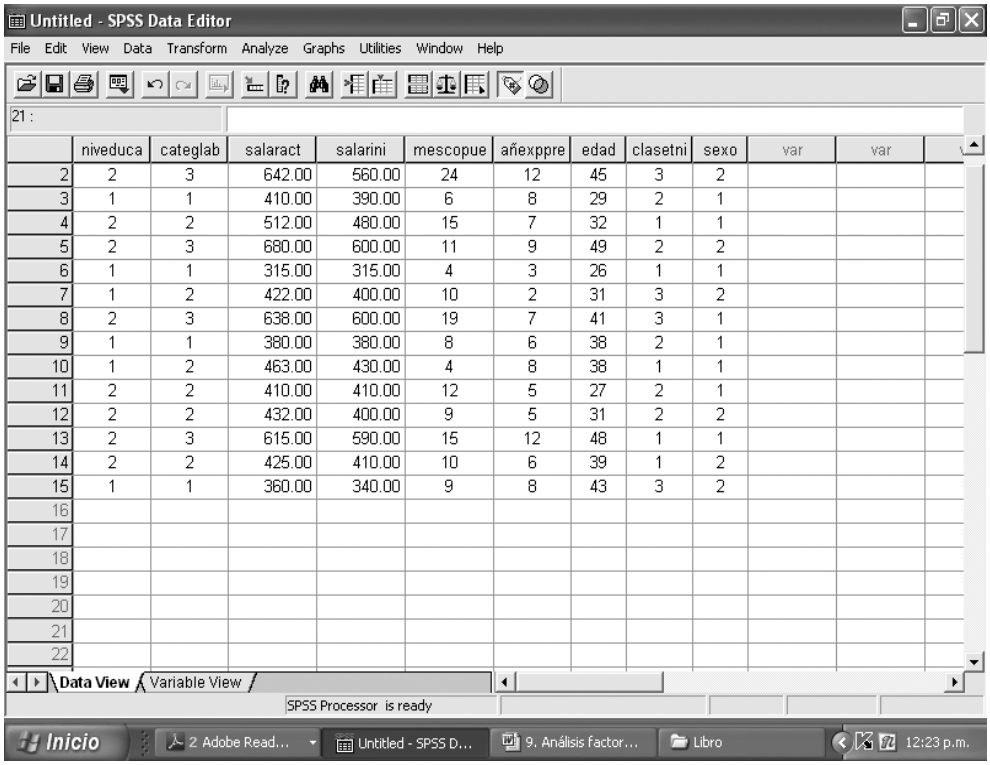
2: mestizo

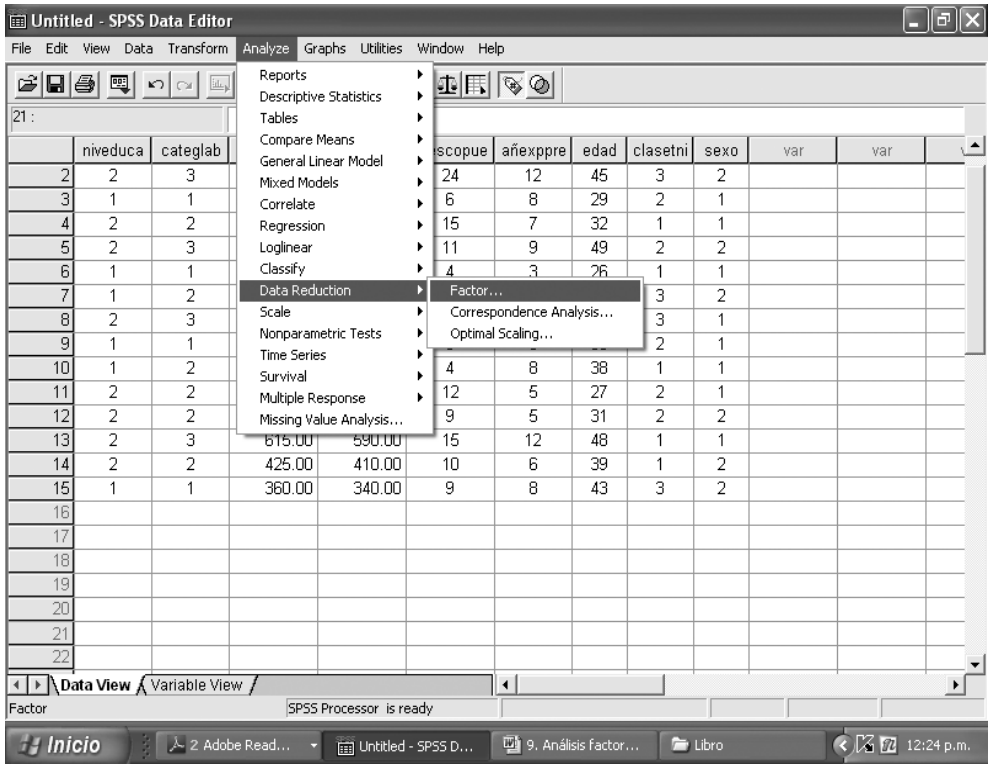
3: negro

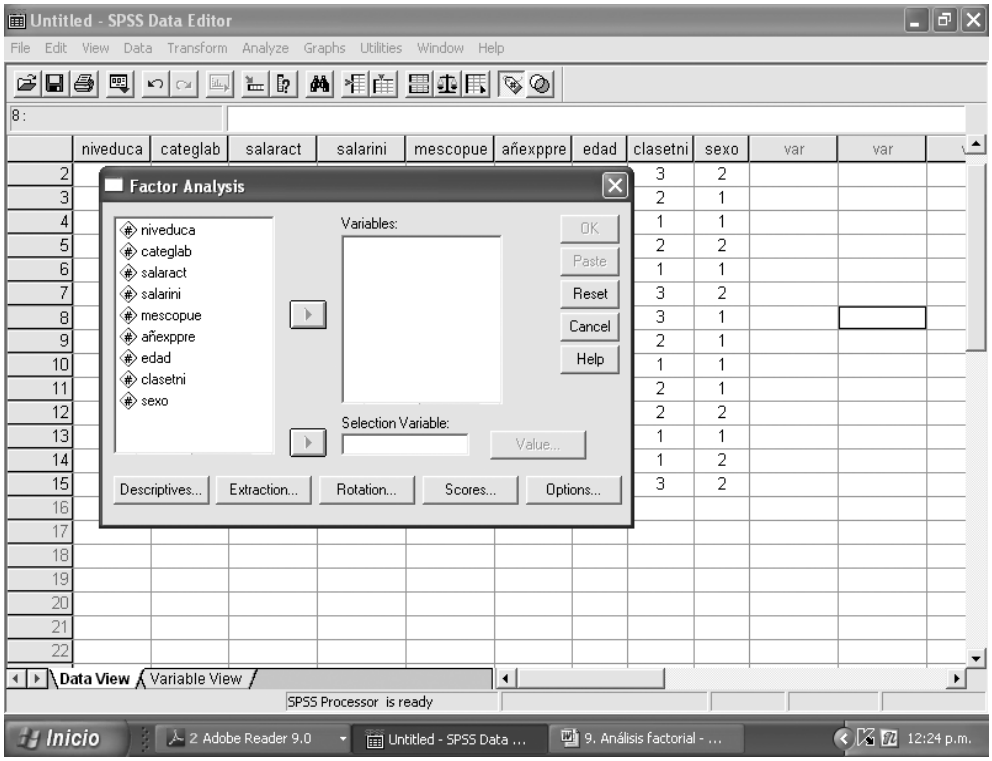
Sexo:

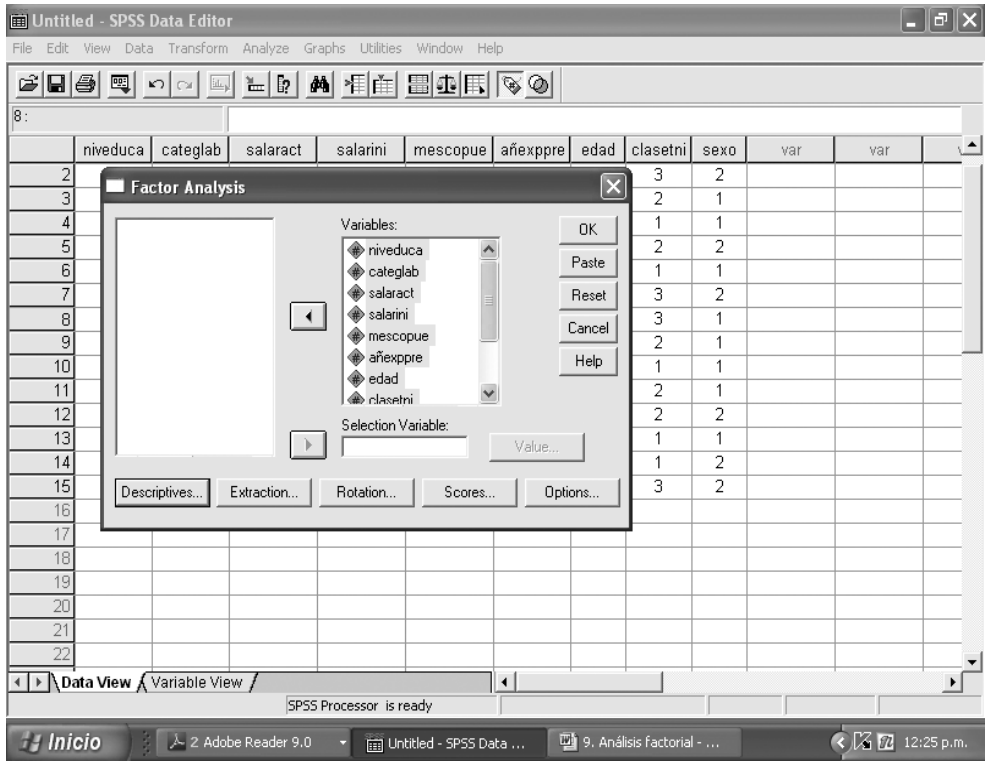
1: femenino

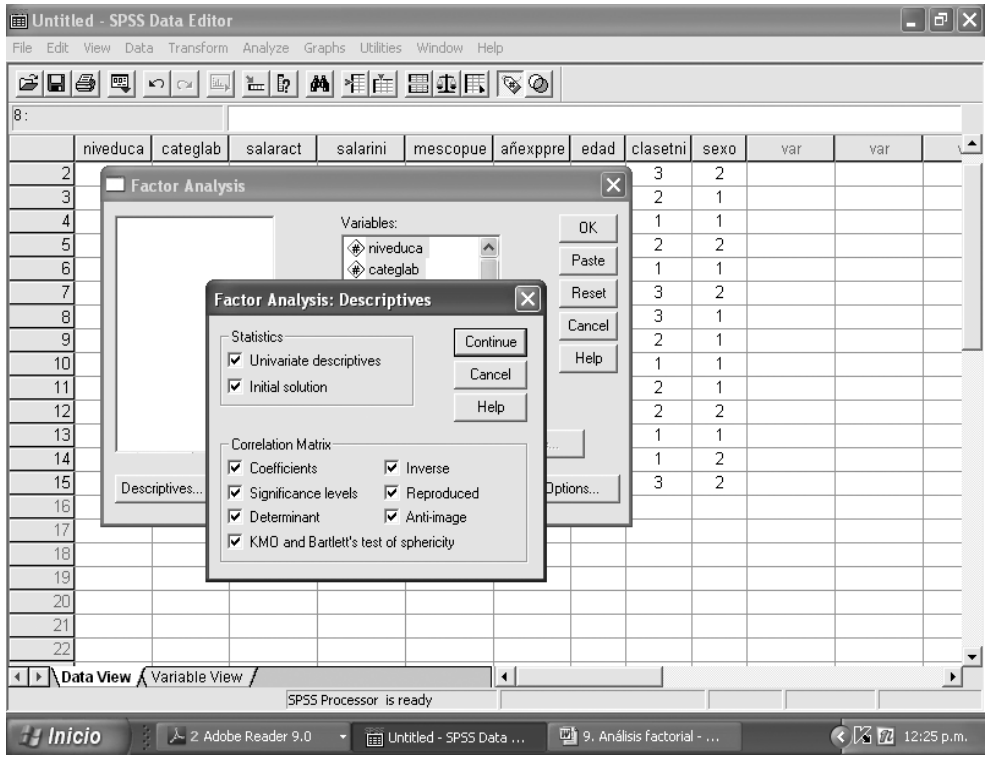
2: masculino

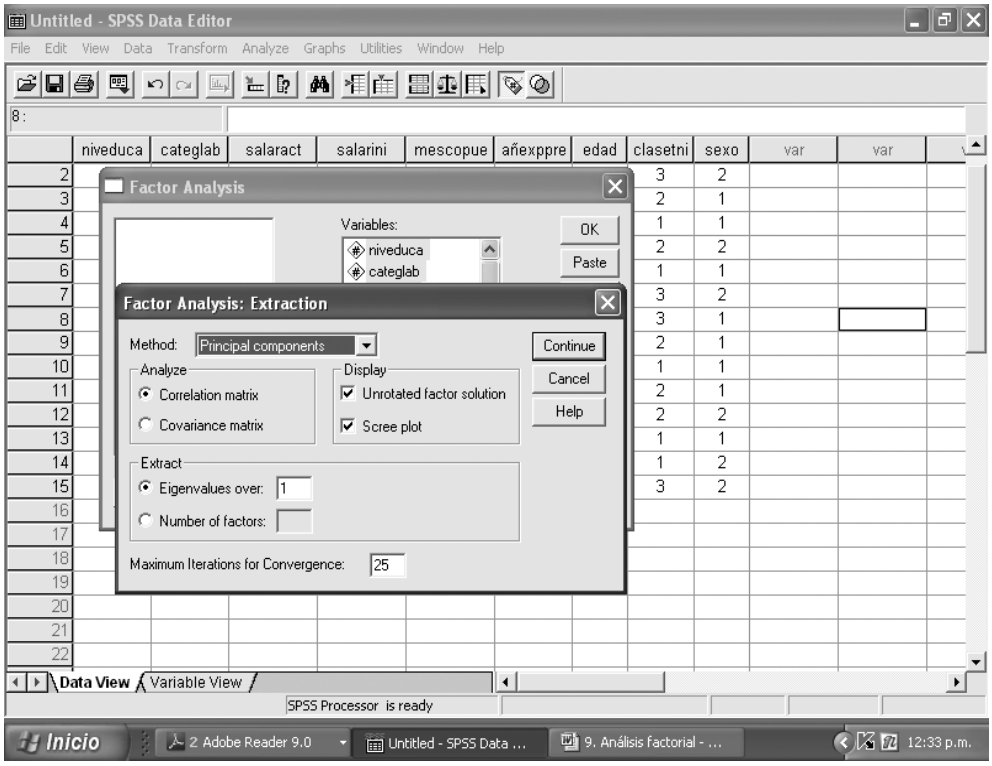


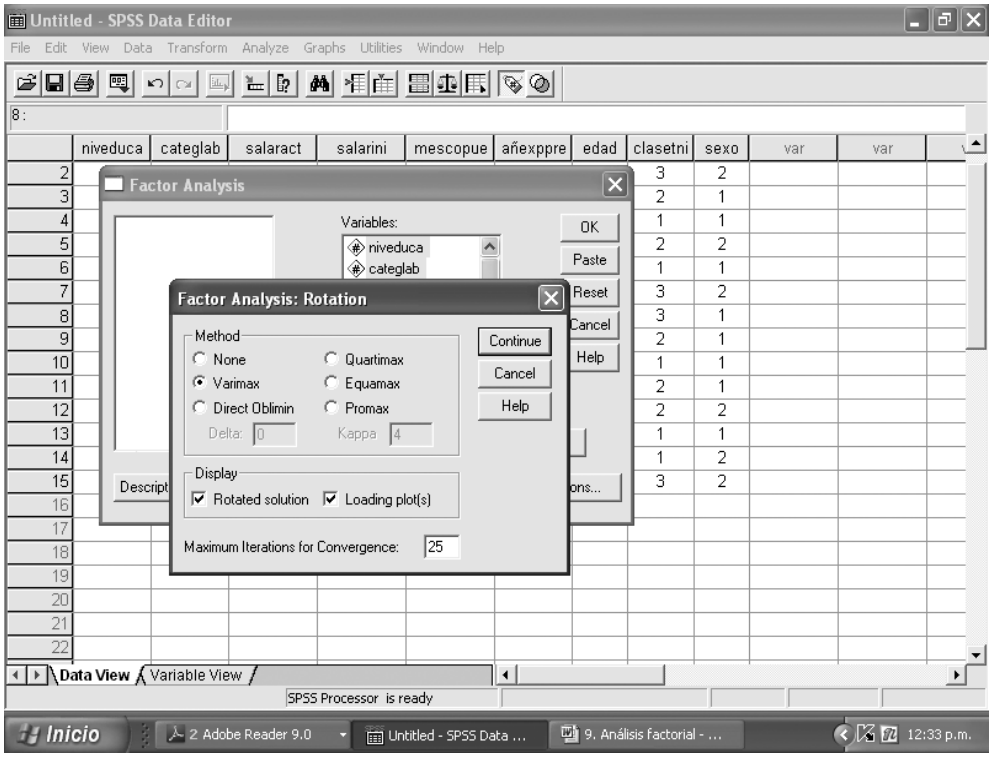

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Factor...
Correspondence Analysis...
Optimal Scaling...
niveduca categlab escopue añexppre edad clasetni sexo var var
Data View Variable View
Factor
SPSS Processor is ready
Inicio
2 Adobe Read...
Untitled - SPSS D...
9. Análisis factor...
Libro
12:24 p.m.

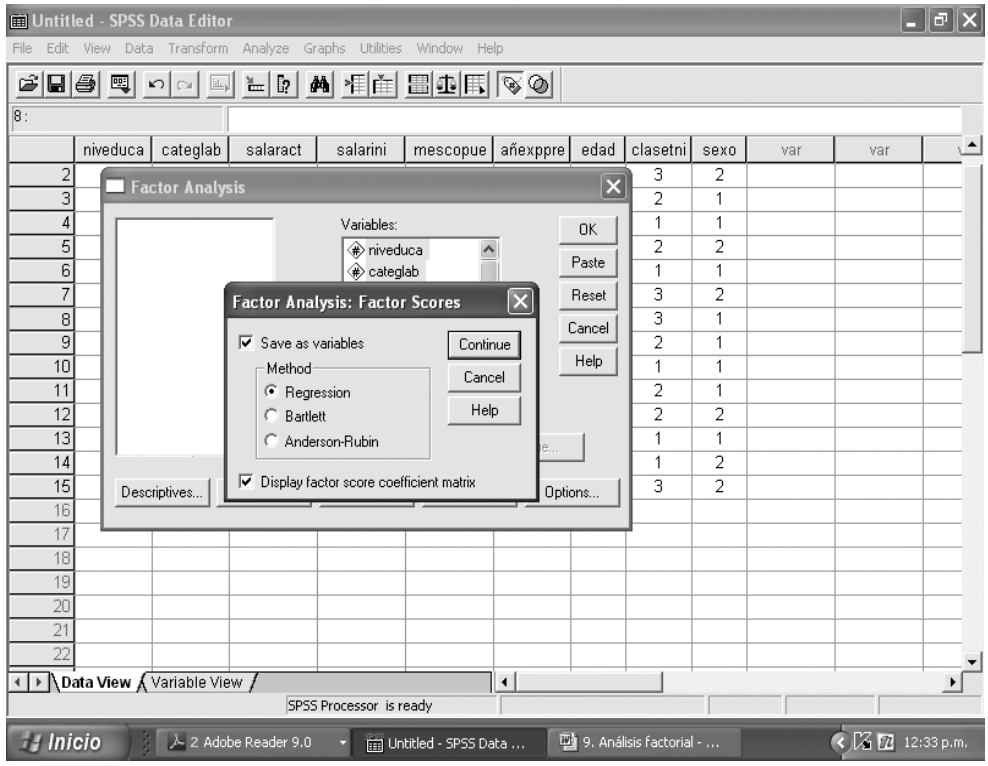

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
niveduca categlab salaract salarini mescopue añexppre edad clasetni sexo var var
Factor Analysis
niveduca
categlab
salaract
salarini
mescopue
añexppre
edad
clasetni
sexo
Variables:
Selection Variable:
Value...
OK
Paste
Reset
Cancel
Help
Descriptives...
Extraction...
Rotation...
Scores...
Options...
Data View Variable View
SPSS Processor is ready
Inicio
2 Adobe Reader 9.0
Untitled - SPSS Data ...
9. Análisis factorial - ...
12:24 p.m.

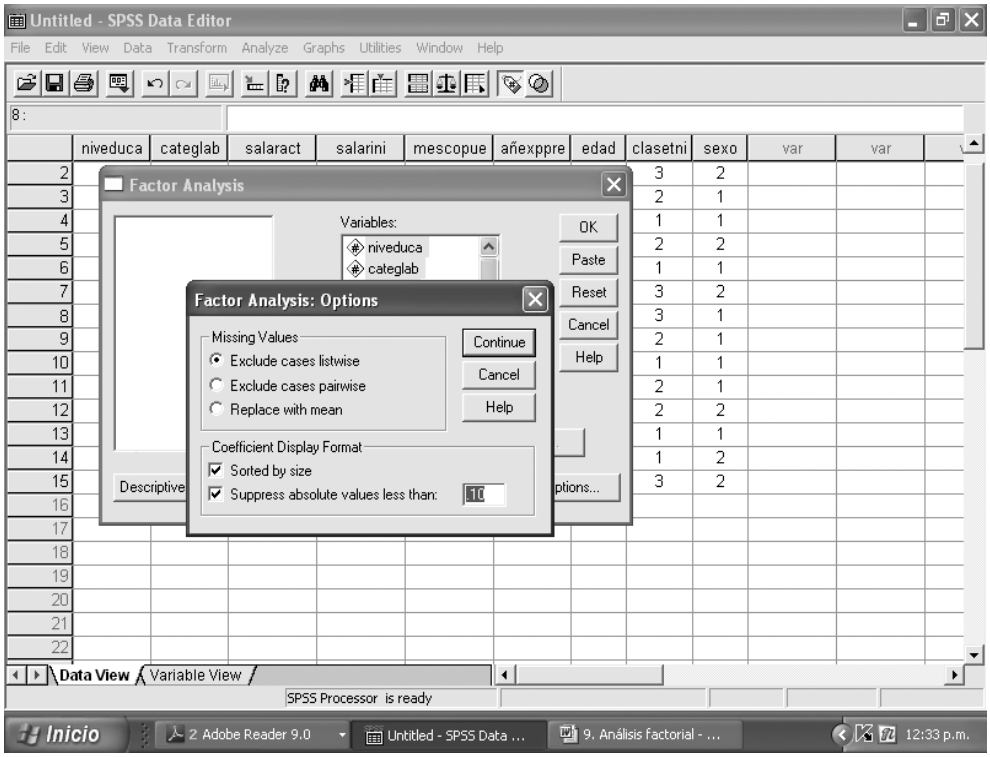

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
niveduca categlab salaract salarini mescopue añexppre edad clasetni sexo var var
Factor Analysis
Variables:
niveduca
categlab
salaract
salarini
mescopue
añexppre
edad
clasetni
OK
Paste
Reset
Cancel
Help
Selection Variable:
Value...
Descriptives...
Extraction...
Rotation...
Scores...
Options...
Data View
Variable View
SPSS Processor is ready
Inicio
2 Adobe Reader 9.0
Untitled - SPSS Data ...
9. Análisis factorial - ...
12:25 p.m.

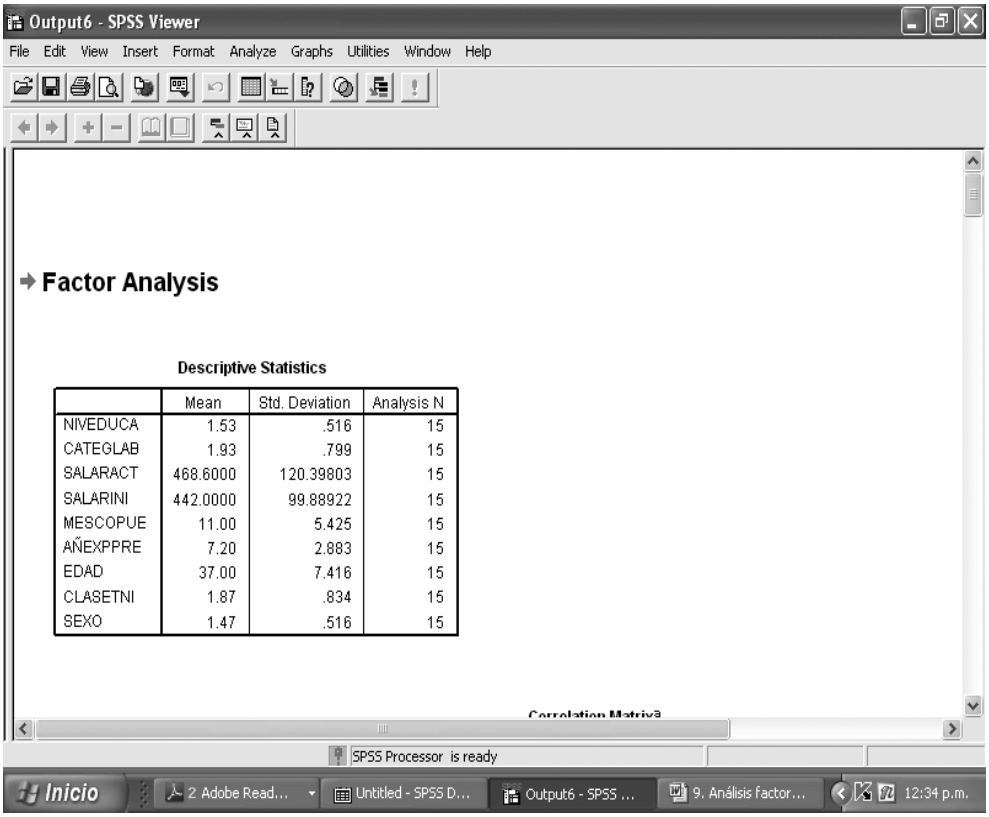

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
niveduca categlab salaract salarini mescopue añexppre edad clasetni sexo var var
Factor Analysis
Variables:
niveduca
categlab
OK
Paste
Reset
Cancel
Help
Descriptives...
Options...
Factor Analysis: Descriptives
Statistics
Univariate descriptives
Initial solution
Continue
Cancel
Help
Correlation Matrix
Coefficients
Significance levels
Determinant
KMO and Bartlett's test of sphericity
Inverse
Reproduced
Anti-image
Data View
Variable View
SPSS Processor is ready
Inicio
2 Adobe Reader 9.0
Untitled - SPSS Data ...
9. Análisis factorial - ...
12:25 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Factor Analysis
Variables:
niveduca
categlab
OK
Paste
Factor Analysis: Extraction
Method: Principal components
Analyze
Correlation matrix
Covariance matrix
Display
Unrotated factor solution
Scree plot
Extract
Eigenvalues over: 1
Number of factors:
Maximum Iterations for Convergence: 25
Continue
Cancel
Help
Data View
Variable View
SPSS Processor is ready
Inicio
2 Adobe Reader 9.0
Untitled - SPSS Data ...
9. Análisis factorial - ...
12:33 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Factor Analysis
Variables:
niveduca
categlab
OK
Paste
Reset
Cancel
Help
Factor Analysis: Rotation
Method
None
Varimax
Direct Oblimin
Quartimax
Equamax
Promax
Delta: 0
Kappa: 4
Continue
Cancel
Help
Display
Rotated solution
Loading plot(s)
Maximum Iterations for Convergence: 25
Data View
Variable View
SPSS Processor is ready
Inicio
2 Adobe Reader 9.0
Untitled - SPSS Data ...
9. Análisis factorial - ...
12:33 p.m.

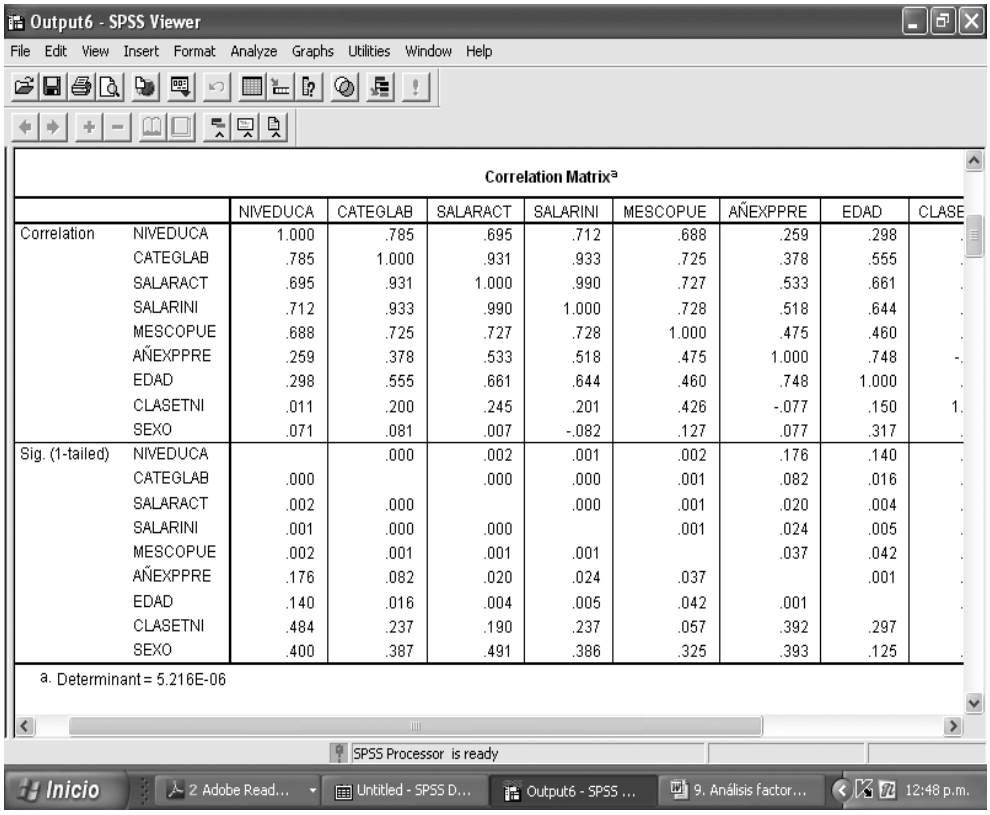

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Factor Analysis
Variables:
niveduca
categlab
OK
Paste
Reset
Cancel
Help
Factor Analysis: Factor Scores
Save as variables
Method
Regression
Bartlett
Anderson-Rubin
Continue
Cancel
Help
Display factor score coefficient matrix
Descriptives...
Options...
Data View
Variable View
SPSS Processor is ready
Inicio
2 Adobe Reader 9.0
Untitled - SPSS Data ...
9. Análisis factorial - ...
12:33 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Factor Analysis
Variables:
niveduca
categlab
OK
Paste
Reset
Cancel
Help
Factor Analysis: Options
Missing Values
Exclude cases listwise
Exclude cases pairwise
Replace with mean
Continue
Cancel
Help
Coefficient Display Format
Sorted by size
Suppress absolute values less than:
Data View
Variable View
SPSS Processor is ready
Inicio
2 Adobe Reader 9.0
Untitled - SPSS Data ...
9. Análisis factorial - ...
12:33 p.m.



	Mean	Std. Deviation	Analysis N
NIVEDUCA	1.53	.516	15
CATEGLAB	1.93	.799	15
SALARACT	468.6000	120.39803	15
SALARINI	442.0000	99.88922	15
MESCOPUE	11.00	5.425	15
AÑEXPPRE	7.20	2.883	15
EDAD	37.00	7.416	15
CLASETNI	1.87	.834	15
SEXO	1.47	.516	15

Véase en la imagen anterior la tabla "*Descriptive Statistics*". En ella se observa el valor promedio y la desviación típica de las observaciones de cada una de las nueve variables de análisis. Como comentario, puede afirmarse que existe un equilibrio entre la cantidad de trabajadores que posee nivel medio y la cantidad que tiene nivel superior. La categoría laboral predominante es la de especialista. El salario promedio actual es de $469.00 mientras que el inicial de $442.00. Los trabajadores llevan en su puesto de trabajo actual, una media de casi un año, mientras que la experiencia previa oscila los siete años. La edad predominante es de 37 años, abundan los empleados mestizos, y las mujeres.


Output6 - SPSS Viewer

File Edit View Insert Format Analyze Graphs Utilities Window Help

Correlation Matrix[a]

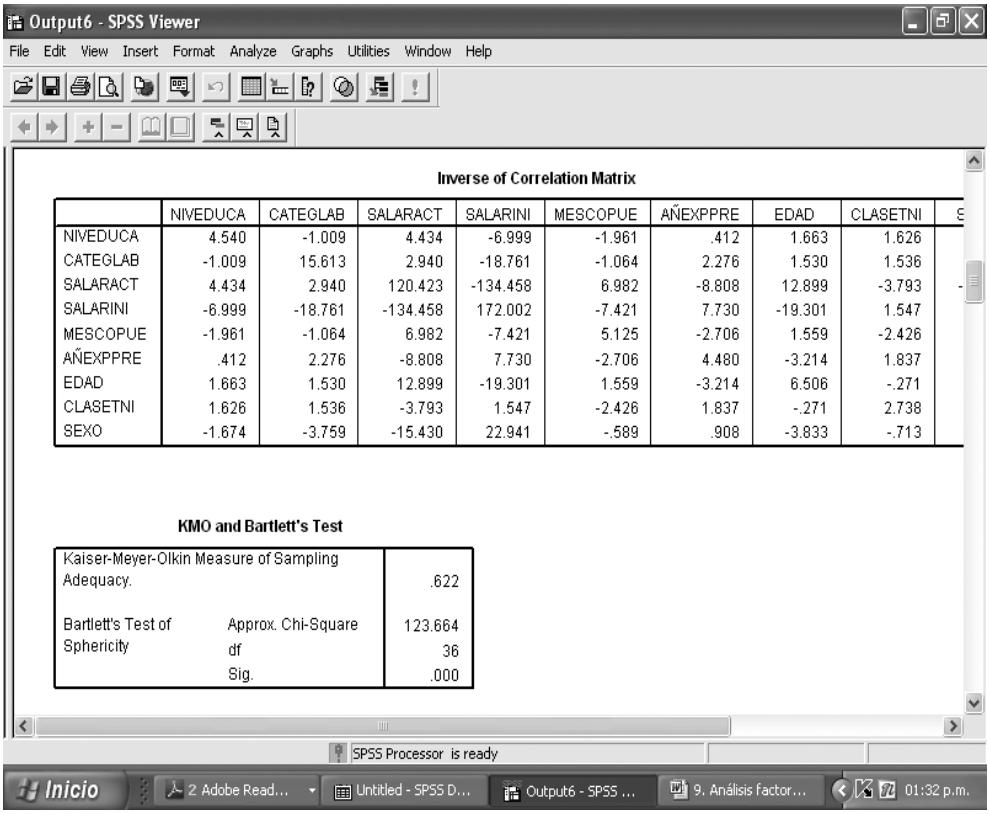
		NIVEDUCA	CATEGLAB	SALARACT	SALARINI	MESCOPUE	AÑEXPPRE	EDAD	CLASE
Correlation	NIVEDUCA	1.000	.785	.695	.712	.688	.259	.298	.
	CATEGLAB	.785	1.000	.931	.933	.725	.378	.555	.
	SALARACT	.695	.931	1.000	.990	.727	.533	.661	.
	SALARINI	.712	.933	.990	1.000	.728	.518	.644	.
	MESCOPUE	.688	.725	.727	.728	1.000	.475	.460	.
	AÑEXPPRE	.259	.378	.533	.518	.475	1.000	.748	-.
	EDAD	.298	.555	.661	.644	.460	.748	1.000	.
	CLASETNI	.011	.200	.245	.201	.426	-.077	.150	1.
	SEXO	.071	.081	.007	-.082	.127	.077	.317	.
Sig. (1-tailed)	NIVEDUCA		.000	.002	.001	.002	.176	.140	.
	CATEGLAB	.000		.000	.000	.001	.082	.016	.
	SALARACT	.002	.000		.000	.001	.020	.004	.
	SALARINI	.001	.000	.000		.001	.024	.005	.
	MESCOPUE	.002	.001	.001	.001		.037	.042	.
	AÑEXPPRE	.176	.082	.020	.024	.037		.001	.
	EDAD	.140	.016	.004	.005	.042	.001		.
	CLASETNI	.484	.237	.190	.237	.057	.392	.297	
	SEXO	.400	.387	.491	.386	.325	.393	.125	.

a. Determinant = 5.216E-06

SPSS Processor is ready

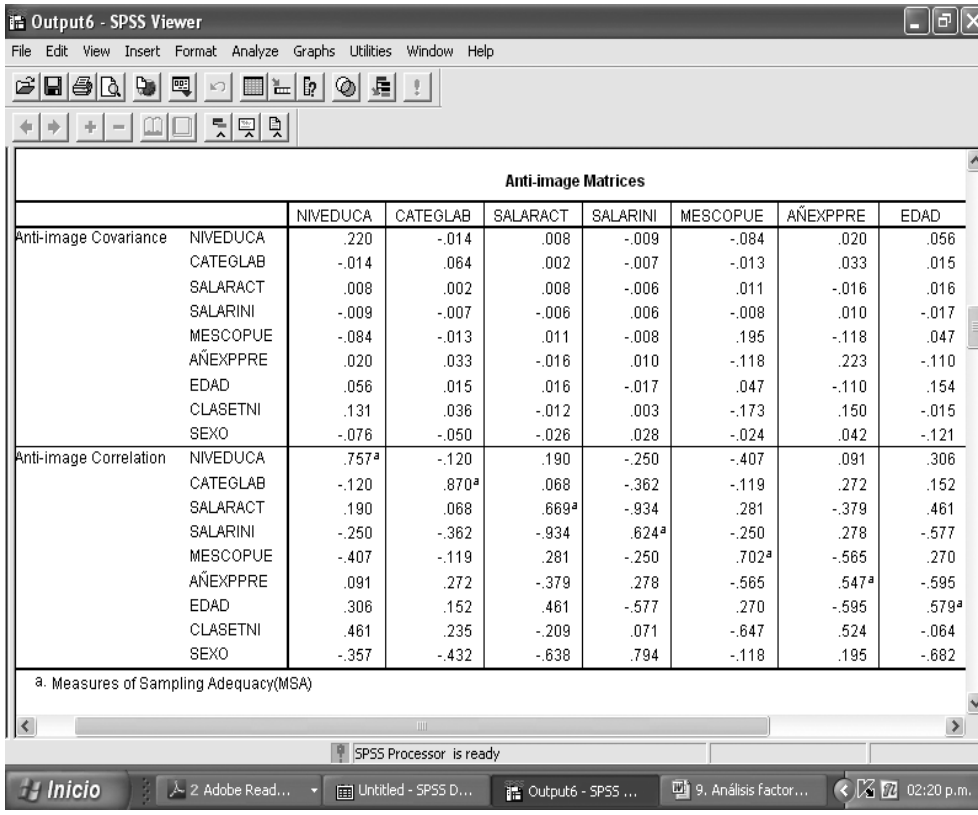
Inicio | 2 Adobe Read... | Untitled - SPSS D... | Output6 - SPSS ... | 9. Análisis factor... | 12:48 p.m.

En la imagen anterior, aparece la tabla "*Correlation Matrix*" donde se reflejan los coeficientes de correlación de Pearson entre cada par de variables, así como la significación de cada correlación. Véase que la mayoría de los coeficientes son elevados y los niveles de significación pequeños. Esto se corrobora con el valor del determinante que es muy pequeño bien cercano a cero, indicando que el análisis factorial es una técnica eficiente en este caso. Puede afirmarse, por ejemplo, que la correlación más elevada se produce entre las variables "salario actual" y "salario inicial" con un valor de 0.99, mientras que la mínima, entre "salario actual" y "sexo" equivalente a 0.007.



En la imagen anterior, aparece una primera tabla "*Inverse of Correlation Matrix*" donde se muestra la matriz inversa de la matriz de correlaciones vista previamente. Ha sido posible obtenerla porque el determinante tuvo un valor distinto de cero. Esto permite continuar el análisis con el método de extracción de factores elegido inicialmente en este ejemplo (componentes principales), pero en caso de querer intentar otro método, sí sería posible también.

La segunda tabla "*KMO and Bartlett´s Test*" muestra un valor de KMO igual a 0.62 que se considera regular. No obstante, como se halla por encima de 0.60, indica que se puede proseguir el análisis factorial con los datos del ejemplo actual. Igualmente, la prueba de esfericidad de Bartlett, muestra un valor de probabilidad igual a 0.000 que es mucho más pequeño que 0.05. Quiere decir que se cumple la región crítica y, por tanto, se rechaza la hipótesis nula, reflejando que la matriz de correlaciones, es una matriz de no identidad. Se corrobora nuevamente que el modelo factorial es adecuado con los datos que se poseen.


Output6 - SPSS Viewer

File Edit View Insert Format Analyze Graphs Utilities Window Help

Anti-image Matrices

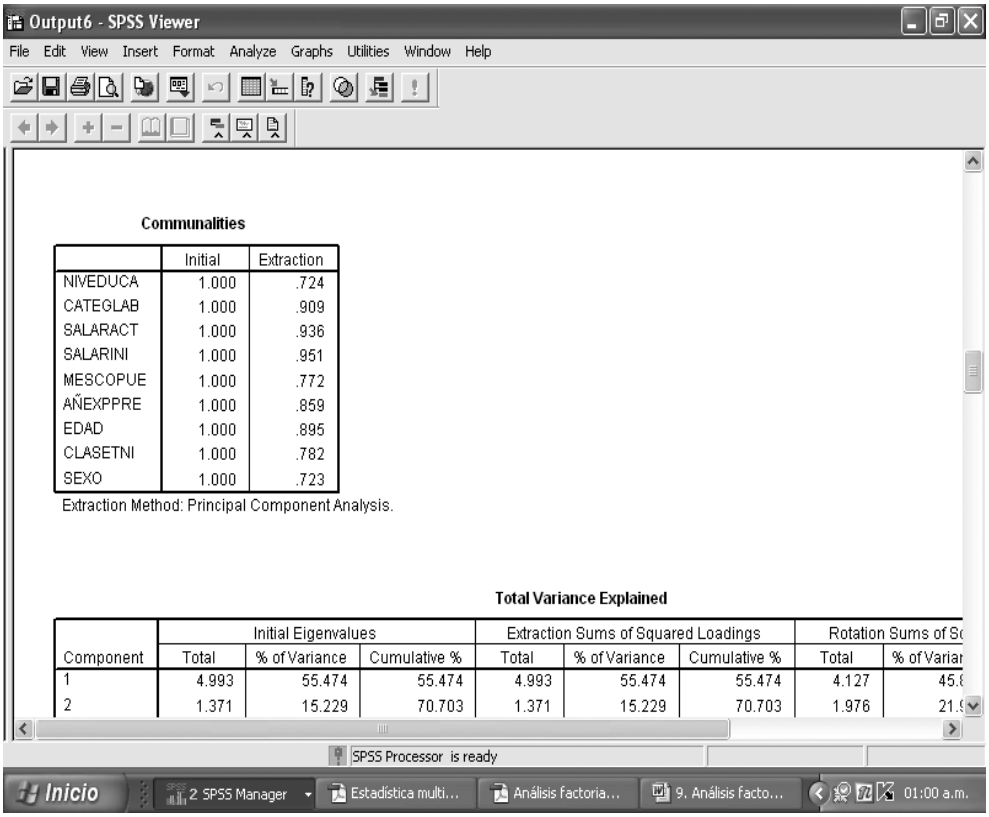
		NIVEDUCA	CATEGLAB	SALARACT	SALARINI	MESCOPUE	AÑEXPPRE	EDAD
Anti-image Covariance	NIVEDUCA	.220	-.014	.008	-.009	-.084	.020	.056
	CATEGLAB	-.014	.064	.002	-.007	-.013	.033	.015
	SALARACT	.008	.002	.008	-.006	.011	-.016	.016
	SALARINI	-.009	-.007	-.006	.006	-.008	.010	-.017
	MESCOPUE	-.084	-.013	.011	-.008	.195	-.118	.047
	AÑEXPPRE	.020	.033	-.016	.010	-.118	.223	-.110
	EDAD	.056	.015	.016	-.017	.047	-.110	.154
	CLASETNI	.131	.036	-.012	.003	-.173	.150	-.015
	SEXO	-.076	-.050	-.026	.028	-.024	.042	-.121
Anti-image Correlation	NIVEDUCA	.757[a]	-.120	.190	-.250	-.407	.091	.306
	CATEGLAB	-.120	.870[a]	.068	-.362	-.119	.272	.152
	SALARACT	.190	.068	.669[a]	-.934	.281	-.379	.461
	SALARINI	-.250	-.362	-.934	.624[a]	-.250	.278	-.577
	MESCOPUE	-.407	-.119	.281	-.250	.702[a]	-.565	.270
	AÑEXPPRE	.091	.272	-.379	.278	-.565	.547[a]	-.595
	EDAD	.306	.152	.461	-.577	.270	-.595	.579[a]
	CLASETNI	.461	.235	-.209	.071	-.647	.524	-.064
	SEXO	-.357	-.432	-.638	.794	-.118	.195	-.682

a. Measures of Sampling Adequacy(MSA)

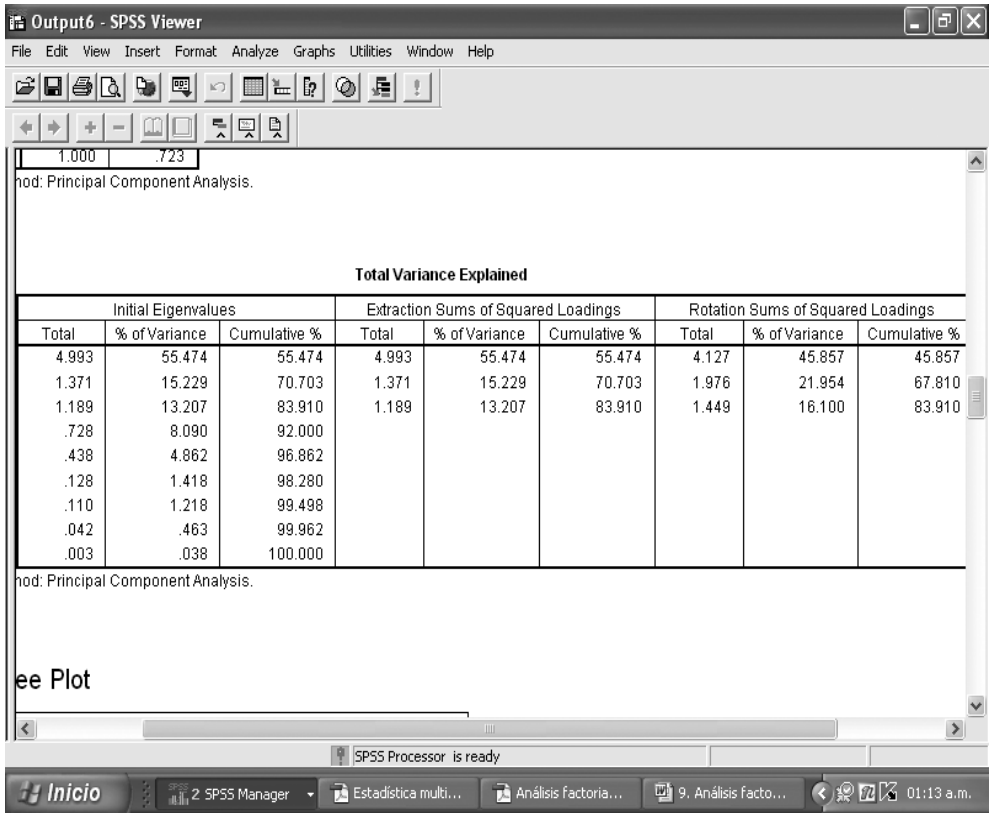
SPSS Processor is ready

Inicio | 2 Adobe Read... | Untitled - SPSS D... | Output6 - SPSS ... | 9. Análisis factor... | 02:20 p.m.

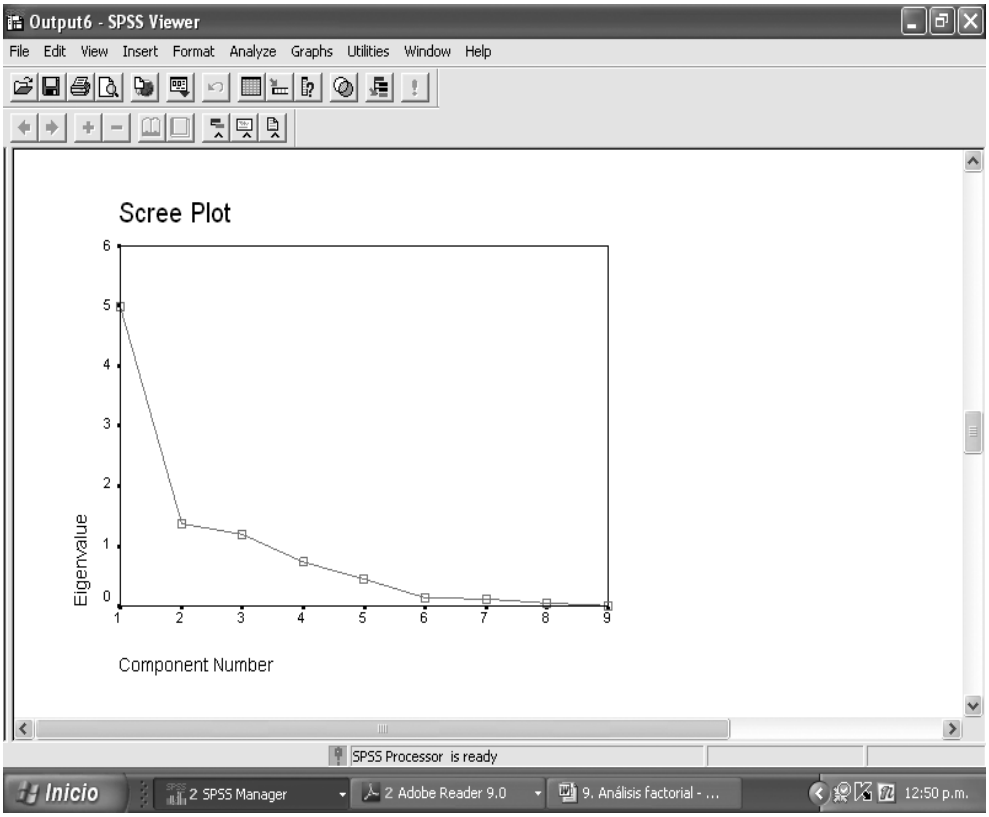
En la imagen anterior, se muestra la tabla "*Anti-image Matrices*", donde se observa que los valores de los coeficientes, en su mayoría, son pequeños. En la diagonal principal de la matriz de correlación anti-imagen, la mayoría de los coeficientes son elevados cercanos a 1 (por ejemplo: 0.757, 0.870, 0.669, 0.624, etc.), siendo de un significado similar al estadístico KMO. Por tanto, el modelo factorial es el adecuado con los datos que se tienen.



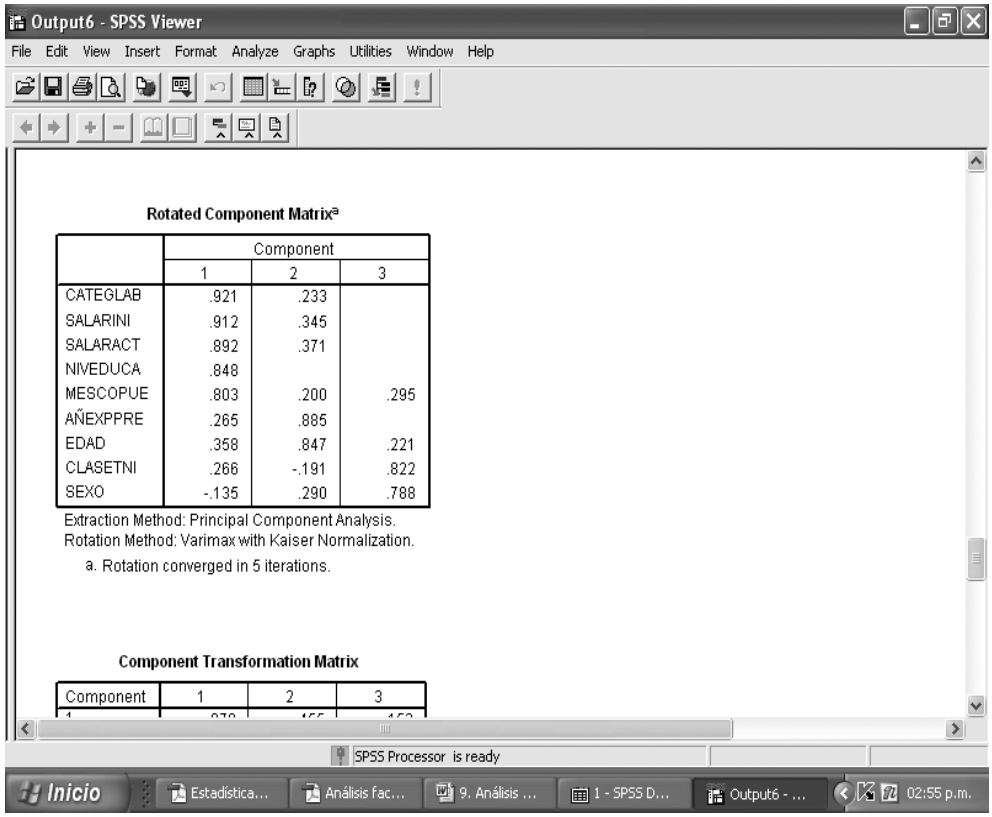
En la imagen anterior se observa la tabla "*Communalities*", donde se muestra que la estimación inicial de la comunalidad de cada variable, es elevada igual a 1. Después de finalizada la extracción, la comunalidad de cada variable sigue siendo elevada, por lo cual todas continúan incluidas para proseguir con el análisis factorial.



En la imagen anterior, se muestra la tabla "*Total Variance Explained*" donde se observa tres autovalores (eigenvalues) superiores a 1 (4.993, 1.371, 1.189) indicando que han sido extraídos tres factores. Véase que estos últimos incluidos en el modelo, son capaces de explicar el 83.91% de la variabilidad total, lo cual puede interpretarse como un porcentaje bastante elevado y, por tanto, muy aceptable.



En la imagen anterior, se muestra el gráfico de sedimentación de Cattell "*Scree Plot*" donde en el eje de las ordenadas, se hallan los autovalores de cada variable, y en el eje de las abscisas, la cantidad de factores. Véase que a partir del tercer autovalor, el resto posee valores inferiores a 1, y que la pendiente que todos ellos forman en conjunto, deja de ser significativa (pierde inclinación) a partir del tercer autovalor en adelante (observando de izquierda a derecha el gráfico). Esto corrobora visualmente, que deben ser tomados en cuenta tres factores para reducir los datos de las nueve variables iniciales.



En la imagen anterior, se muestra la tabla "*Rotated Component Matrix*", donde se observa cuáles variables corresponden a cada uno de los tres factores. Se señala que este resultado, es la consecuencia de haber seleccionado como método de extracción de factores, el de componentes principales, y como método de rotación, el varimax. Véase que según los valores de los coeficientes:

En el primer factor se hallan las variables:

- categoría laboral (0.921)
- salario inicial (0.912)
- salario actual (0.892)
- nivel educativo (0.848)
- meses desde el comienzo en el puesto (0.803)

En el segundo factor:

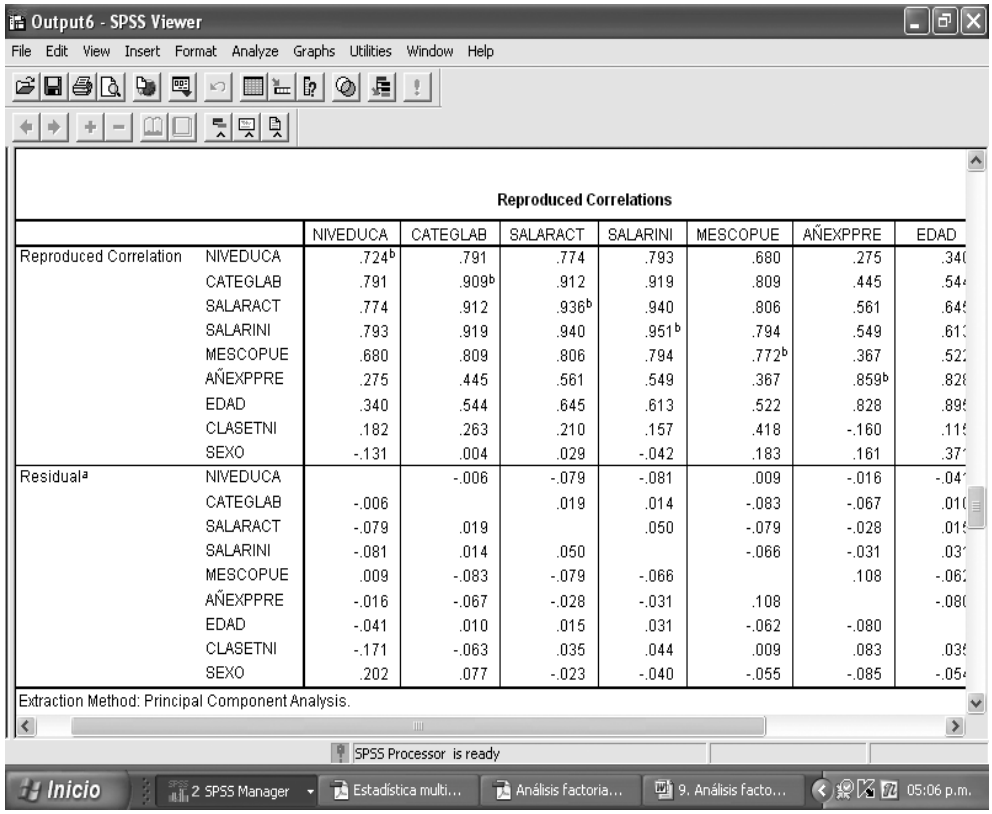
- años de experiencia previa (0.885)
- edad (0.847)

En el tercer factor:

- clasificación étnica (0.822)
- sexo (0.788)

Es recomendable asignarle un nombre a cada factor, y en este ejemplo pudiera ser:

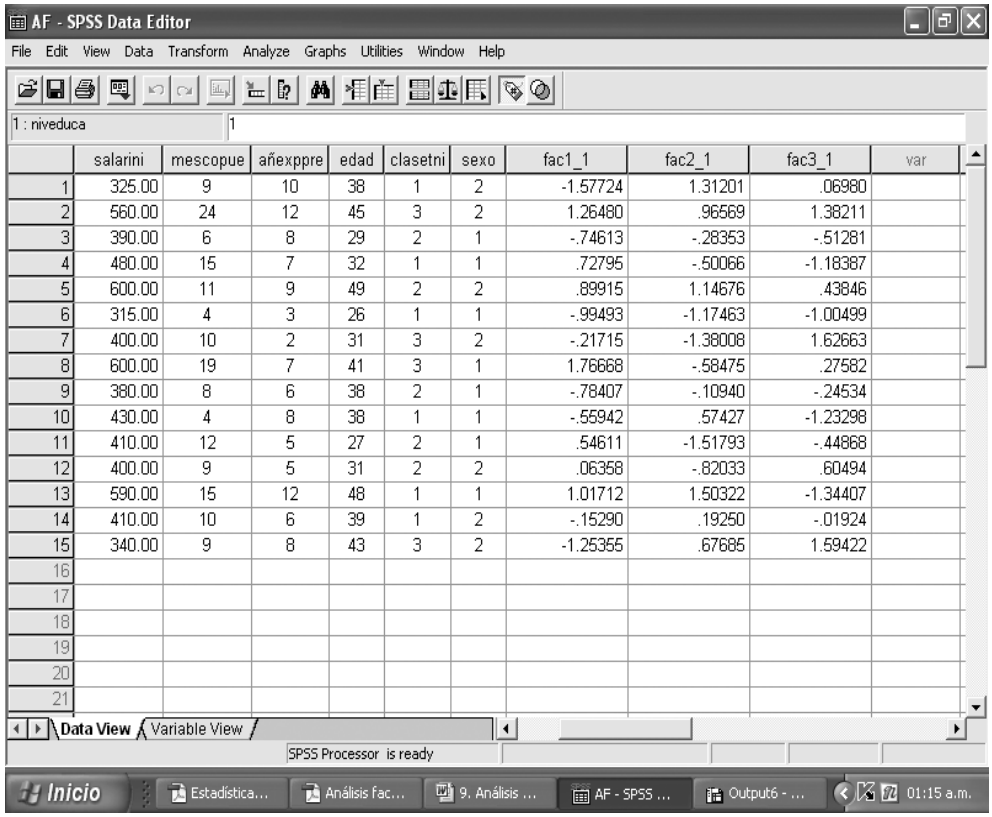
- Primer factor: promoción laboral
- Segundo factor: veteranía laboral
- Tercer factor: identificación física



		NIVEDUCA	CATEGLAB	SALARACT	SALARINI	MESCOPUE	AÑEXPPRE	EDAD
Reproduced Correlation	NIVEDUCA	.724[b]	.791	.774	.793	.680	.275	.340
	CATEGLAB	.791	.909[b]	.912	.919	.809	.445	.544
	SALARACT	.774	.912	.936[b]	.940	.806	.561	.645
	SALARINI	.793	.919	.940	.951[b]	.794	.549	.613
	MESCOPUE	.680	.809	.806	.794	.772[b]	.367	.522
	AÑEXPPRE	.275	.445	.561	.549	.367	.859[b]	.828
	EDAD	.340	.544	.645	.613	.522	.828	.895
	CLASETNI	.182	.263	.210	.157	.418	-.160	.119
	SEXO	-.131	.004	.029	-.042	.183	.161	.371
Residual[a]	NIVEDUCA		-.006	-.079	-.081	.009	-.016	-.041
	CATEGLAB	-.006		.019	.014	-.083	-.067	.010
	SALARACT	-.079	.019		.050	-.079	-.028	.019
	SALARINI	-.081	.014	.050		-.066	-.031	.031
	MESCOPUE	.009	-.083	-.079	-.066		.108	-.062
	AÑEXPPRE	-.016	-.067	-.028	-.031	.108		-.080
	EDAD	-.041	.010	.015	.031	-.062	-.080	
	CLASETNI	-.171	-.063	.035	.044	.009	.083	.038
	SEXO	.202	.077	-.023	-.040	-.055	-.085	-.054

Extraction Method: Principal Component Analysis.

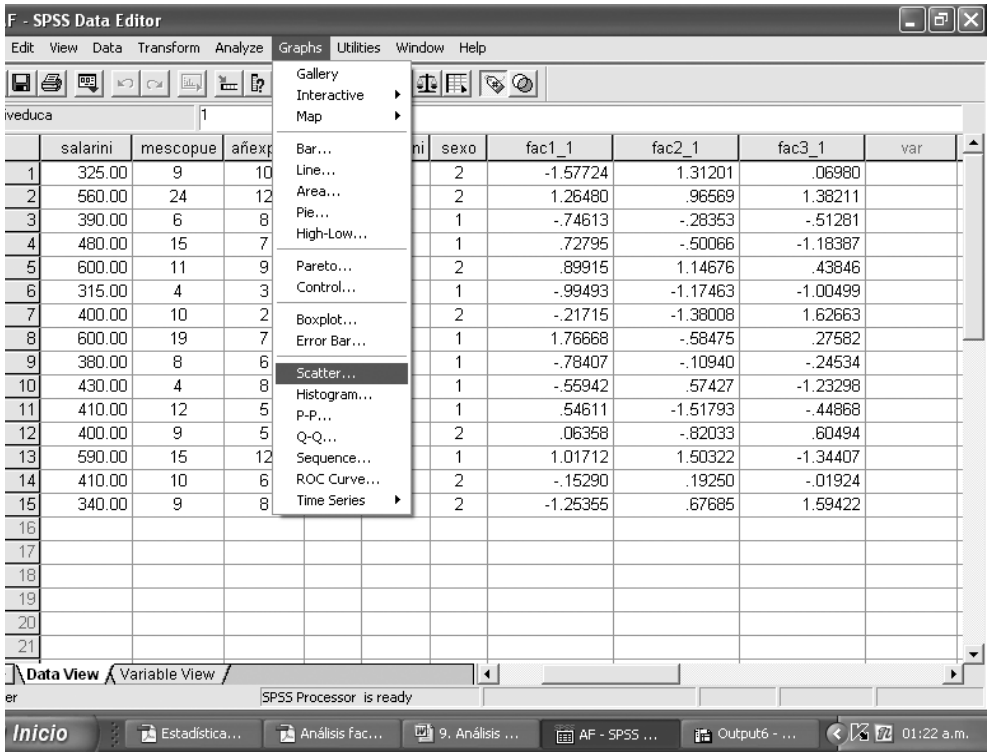
En la imagen anterior, se muestra la tabla "*Reproduced Correlations*" donde se observa la matriz de correlaciones reproducidas y la matriz residual. Como en la matriz residual todos los coeficientes están muy próximos a 0, esto indica que el análisis factorial ha sido fructífero.

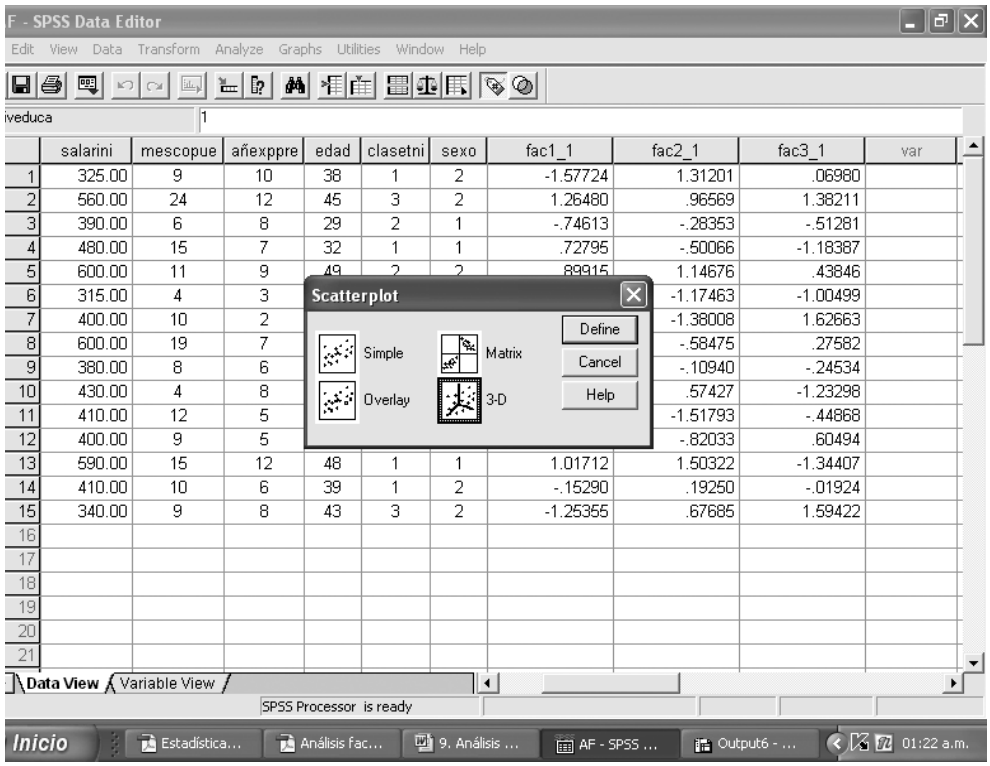


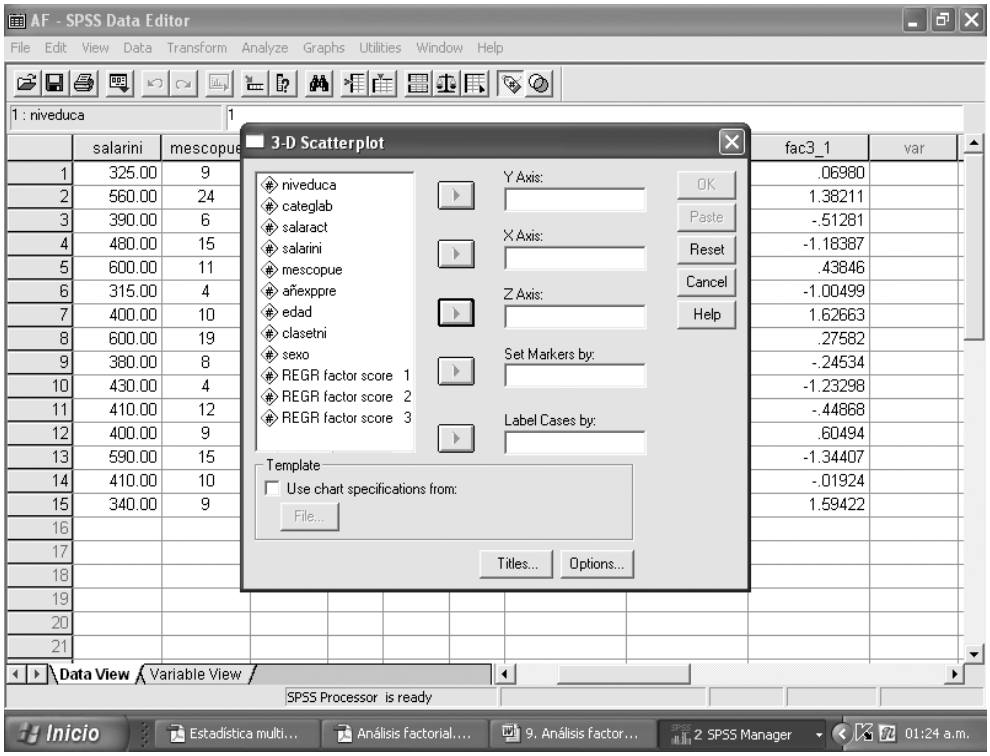
	salarini	mescopue	añexppre	edad	clasetni	sexo	fac1_1	fac2_1	fac3_1	var
1	325.00	9	10	38	1	2	-1.57724	1.31201	.06980	
2	560.00	24	12	45	3	2	1.26480	.96569	1.38211	
3	390.00	6	8	29	2	1	-.74613	-.28353	-.51281	
4	480.00	15	7	32	1	1	.72795	-.50066	-1.18387	
5	600.00	11	9	49	2	2	.89915	1.14676	.43846	
6	315.00	4	3	26	1	1	-.99493	-1.17463	-1.00499	
7	400.00	10	2	31	3	2	-.21715	-1.38008	1.62663	
8	600.00	19	7	41	3	1	1.76668	-.58475	.27582	
9	380.00	8	6	38	2	1	-.78407	-.10940	-.24534	
10	430.00	4	8	38	1	1	-.55942	.57427	-1.23298	
11	410.00	12	5	27	2	1	.54611	-1.51793	-.44868	
12	400.00	9	5	31	2	2	.06358	-.82033	.60494	
13	590.00	15	12	48	1	1	1.01712	1.50322	-1.34407	
14	410.00	10	6	39	1	2	-.15290	.19250	-.01924	
15	340.00	9	8	43	3	2	-1.25355	.67685	1.59422	
16										
17										
18										
19										
20										
21										

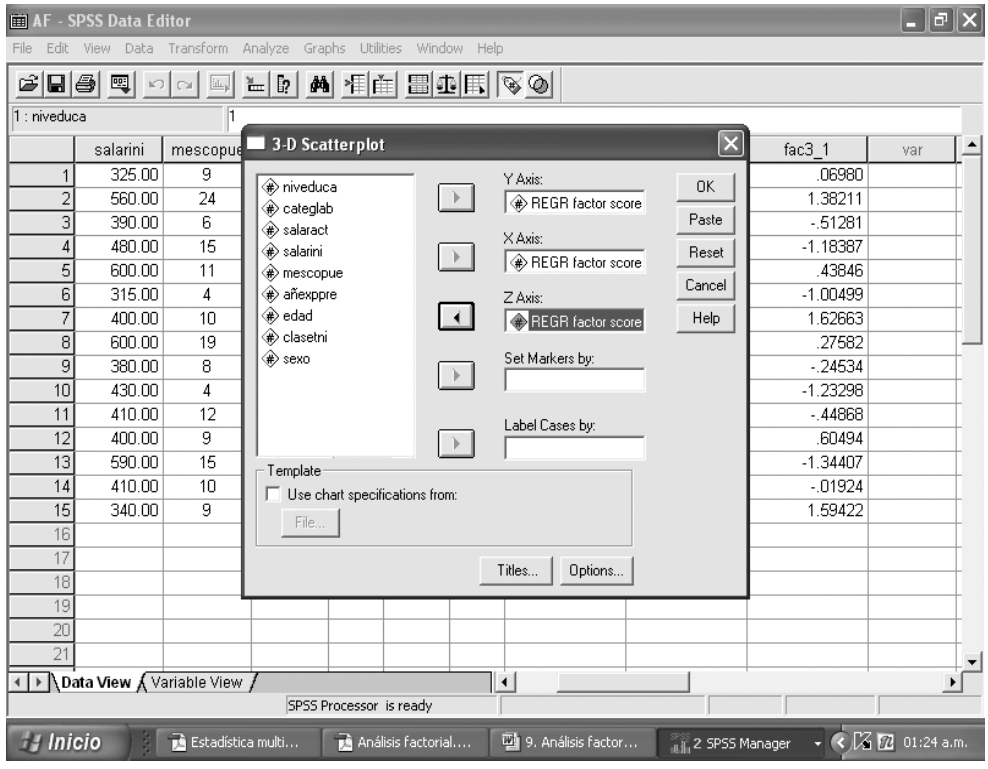
En la imagen anterior, se muestra la base de datos original colocada en el SPSS. El programa ha añadido tres nuevas columnas (fac1_1, fac2_1, fac3_1) que representan las puntuaciones factoriales de cada sujeto en cada uno de los factores extraídos. Esto sirve para representar gráficamente, la posición de cada sujeto de la muestra en el hiperplano.

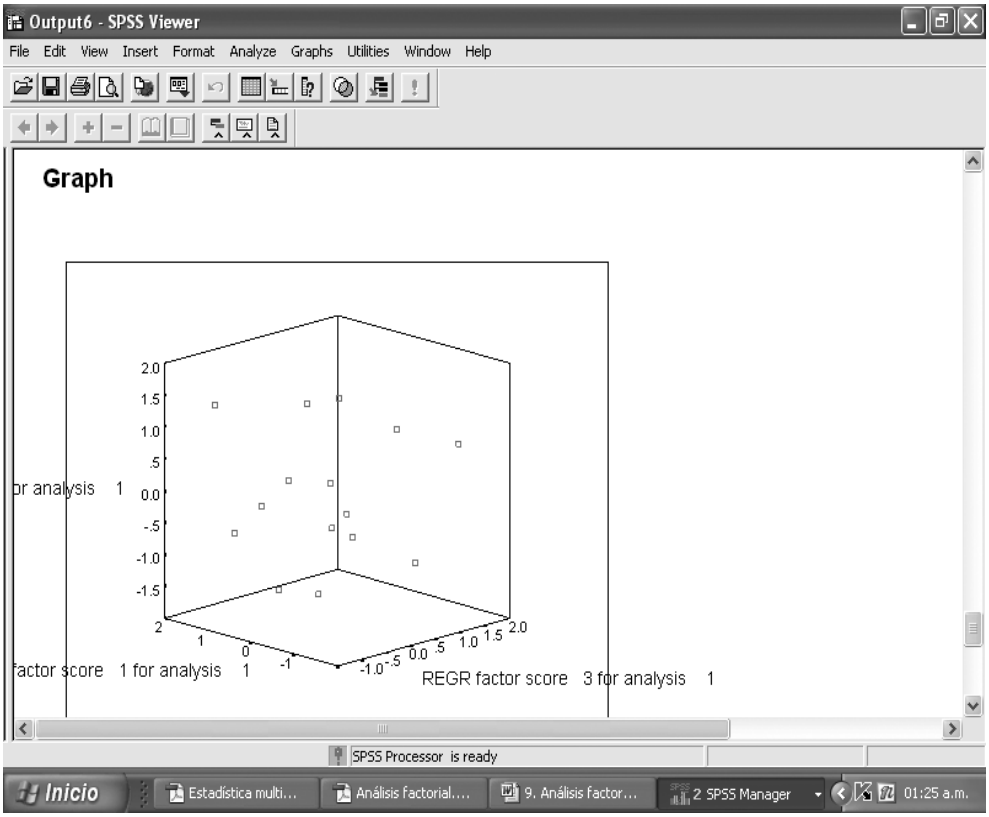
Para obtener esta gráfica, sería:


F - SPSS Data Editor
Edit View Data Transform Analyze Graphs Utilities Window Help
Gallery
Interactive
Map
Bar...
Line...
Area...
Pie...
High-Low...
Pareto...
Control...
Boxplot...
Error Bar...
Scatter...
Histogram...
P-P...
Q-Q...
Sequence...
ROC Curve...
Time Series
salarini mescopue añexp sexo fac1_1 fac2_1 fac3_1 var
1 325.00 9 10 2 -1.57724 1.31201 .06980
2 560.00 24 12 2 1.26480 .96569 1.38211
3 390.00 6 8 1 -.74613 -.28353 -.51281
4 480.00 15 7 1 .72795 -.50066 -1.18387
5 600.00 11 9 2 .89915 1.14676 .43846
6 315.00 4 3 1 -.99493 -1.17463 -1.00499
7 400.00 10 2 2 -.21715 -1.38008 1.62663
8 600.00 19 7 1 1.76668 -.58475 .27582
9 380.00 8 6 1 -.78407 -.10940 -.24534
10 430.00 4 8 1 -.55942 .57427 -1.23298
11 410.00 12 5 1 .54611 -1.51793 -.44868
12 400.00 9 5 2 .06358 -.82033 .60494
13 590.00 15 12 1 1.01712 1.50322 -1.34407
14 410.00 10 6 2 -.15290 .19250 -.01924
15 340.00 9 8 2 -1.25355 .67685 1.59422
Data View Variable View
SPSS Processor is ready
Inicio
Estadística...
Análisis fac...
9. Análisis ...
AF - SPSS ...
Output6 - ...
01:22 a.m.


F - SPSS Data Editor
Edit View Data Transform Analyze Graphs Utilities Window Help
salarini mescopue añexppre edad clasetni sexo fac1_1 fac2_1 fac3_1 var
1 325.00 9 10 38 1 2 -1.57724 1.31201 .06980
2 560.00 24 12 45 3 2 1.26480 .96569 1.38211
3 390.00 6 8 29 2 1 -.74613 -.28353 -.51281
4 480.00 15 7 32 1 1 .72795 -.50066 -1.18387
5 600.00 11 9 49 2 2 .89915 1.14676 .43846
6 315.00 4 3 -1.17463 -1.00499
7 400.00 10 2 -1.38008 1.62663
8 600.00 19 7 -.58475 .27582
9 380.00 8 6 -.10940 -.24534
10 430.00 4 8 .57427 -1.23298
11 410.00 12 5 -1.51793 -.44868
12 400.00 9 5 -.82033 .60494
13 590.00 15 12 48 1 1 1.01712 1.50322 -1.34407
14 410.00 10 6 39 1 2 -.15290 .19250 -.01924
15 340.00 9 8 43 3 2 -1.25355 .67685 1.59422
Scatterplot
Simple
Matrix
Overlay
3-D
Define
Cancel
Help
Data View Variable View
SPSS Processor is ready
Inicio
Estadística...
Análisis fac...
9. Análisis ...
AF - SPSS ...
Output6 - ...
01:22 a.m.


AF - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : niveduca
3-D Scatterplot
niveduca
categlab
salaract
salarini
mescopue
añexppre
edad
clasetni
sexo
REGR factor score 1
REGR factor score 2
REGR factor score 3
Y Axis:
X Axis:
Z Axis:
Set Markers by:
Label Cases by:
Template
Use chart specifications from:
File...
Titles...
Options...
OK
Paste
Reset
Cancel
Help
salarini
mescopue
fac3_1
var
Data View
Variable View
SPSS Processor is ready
Inicio
Estadística multi...
Análisis factorial....
9. Análisis factor...
2 SPSS Manager
01:24 a.m.


AF - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : niveduca
3-D Scatterplot
niveduca
categlab
salaract
salarini
mescopue
añexppre
edad
clasetni
sexo
Y Axis:
REGR factor score
X Axis:
REGR factor score
Z Axis:
REGR factor score
Set Markers by:
Label Cases by:
Template
Use chart specifications from:
File...
Titles...
Options...
OK
Paste
Reset
Cancel
Help
salarini
mescopue
fac3_1
var
Data View
Variable View
SPSS Processor is ready
Inicio
Estadística multi...
Análisis factorial....
9. Análisis factor...
2 SPSS Manager
01:24 a.m.



En la imagen anterior, se observa el gráfico que refleja la posición de cada sujeto en el hiperplano. Véase que de los 15 sujetos estudiados, 7 han dado respuestas similares a las variables (cinco) que corresponden al segundo factor (veteranía laboral). Nótese que 1 sujeto está en el límite del segundo (veteranía laboral) y el tercer factor (identificación física). Igualmente, 5 individuos han dado respuestas similares a las variables (dos) que corresponden al tercer factor (identificación física). Por último, sólo 2 sujetos han dado respuestas similares a las variables (dos) que corresponden al primer factor (promoción laboral).

EJERCITACIÓN

En el Hotel X, los miembros del consejo de dirección, han recopilado los datos pertenecientes a 10 trabajadores de la entidad. Los mismos se distribuyen en siete variables que son:

- nivel de conocimiento de la actividad que realiza
- años de experiencia en el puesto de trabajo
- cantidad de cursos de superación vencidos
- evaluación del desempeño
- nivel de productividad laboral
- nivel de motivación
- salario actual

Se desea entonces, bajo un nivel de confiabilidad del 95%, encontrar un número mínimo de dimensiones (factores), con el objetivo de reducir los datos iniciales partiendo de la similitud que existe entre unas y otras variables anteriores. Véase la tabulación de los datos:

Trabajadores	conocimi	añosexpe	cursos	evaludes	producti	motivaci	salaract
1	2	5	6	2	2	2	345.60
2	3	12	13	4	3	3	645.55
3	1	2	4	2	2	3	298.25
4	1	3	5	1	1	2	276.90
5	3	14	11	4	3	2	599.35
6	3	11	12	5	2	2	610.40
7	3	10	15	5	3	3	624.55
8	3	12	18	4	3	3	589.90
9	2	4	8	3	2	3	388.35
10	3	15	10	5	3	2	597.15
	1: bajo 2: medio 3: alto			1: mal 2: regular 3: bien 4: muy bien 5: excelente	1: bajo 2: medio 3: alto	1: bajo 2: medio 3: alto	

SOLUCIÓN

- en la matriz de correlaciones, los coeficientes de correlación son en su mayoría elevados, y la significación de los mismos, pequeña
- el valor del determinante es igual 5.5^{-6}, o sea, extremadamente pequeño pero distinto de cero
- el valor de KMO es igual a 0.67, es decir, regular, pero mayor que 0.60
- la prueba de esfericidad de Bartlett posee un valor de probabilidad igual a 0.000 demostrando que la matriz de correlaciones, es una matriz de no identidad
- en la tabla de comunalidades, se observa que la estimación inicial de la comunalidad de cada variable, es elevada igual a 1, y después de finalizada la extracción, la comunalidad de cada variable sigue siendo elevada, por lo cual todas continúan incluidas para proseguir con el análisis factorial
- en la tabla de varianza total explicada, se observa dos autovalores superiores a 1 (5.273 y 1.182) indicando que han sido extraídos dos factores, los cuales son capaces de explicar el 99.22% de la variabilidad total, que se interpreta como un porcentaje bastante elevado y, por tanto, muy aceptable
- según el gráfico de sedimentación de Cattell, a partir del segundo autovalor, el resto posee valores inferiores a 1, lo cual corrobora visualmente, que deben ser tomados en cuenta dos factores para reducir los datos de las siete variables iniciales
- en la matriz de componentes rotada según el método varimax:
  En el primer factor que podría llamarse "rendimiento laboral", se hallan las variables:
  - salario actual (0.985)
  - nivel de conocimiento de la actividad que realiza (0.983)
  - años de experiencia en el puesto de trabajo (0.956)
  - evaluación del desempeño (0.938)
  - nivel de productividad laboral (0.871)
  - cantidad de cursos de superación vencidos (0.862)

  En el segundo factor "motivación laboral":
  - nivel de motivación (0.995)

Capítulo
9

Análisis discriminante.

9.1. Concepto de análisis discriminante.

El análisis discriminante, es una técnica estadística que permite asignar un individuo a un grupo definido a priori (variable dependiente), en función de una serie de características del mismo, o de las respuestas dadas a una serie de preguntas (variables independientes).

Se diferencia del análisis clúster (que se verá en el próximo capítulo), en que aquí la clasificación de una muestra en una serie de grupos, se realiza a priori, mientras que en el análisis clúster, se hace a posteriori.

9.2. ¿En qué consiste la función discriminante?

El análisis discriminante es una técnica que permite analizar también, cuáles son las variables que contribuyen en mayor grado, a discriminar (distinguir, diferenciar) a los sujetos en los diferentes grupos definidos a priori. Para ello, estas variables que mejor discriminan, se reducen a variables canónicas, que constituyen una combinación lineal de las variables independientes originales. Dicha combinación lineal es lo que se conoce como función discriminante, donde la variable dependiente es la pertenencia a uno u otro grupo. La importancia de esta función radica en su capacidad de distinguir (discriminar), con la mayor precisión posible, a los miembros de uno u otro grupo.

Variables independientes: son las que permiten discriminar a los sujetos en uno

u otro grupo (llamadas también variables de clasificación o discriminantes).

Variable dependiente: variable categórica con tantos valores discretos como grupos.

Una vez encontrada la función discriminante, podrá ser utilizada para clasificar nuevos casos, individuos, etc.

9.3. Algunas puntualizaciones de interés acerca del análisis discriminante.

Coeficiente de correlación canónica cercano a 0: las variables discriminantes no permiten diferenciar bien entre los grupos.
Coeficiente de correlación canónica cercano a 1: las variables discriminantes permiten diferenciar bien entre los grupos.

Autovalor cercano a 0: las variables discriminantes no permiten diferenciar bien entre los grupos.
Autovalor lejano de 0: las variables discriminantes permiten diferenciar bien entre los grupos.

Estadístico Lambda cercano a 0: gran diferencia entre los grupos.
Estadístico Lambda cercano a 1: gran parecido entre los grupos.

Véase un ejemplo.

Ejemplo 1:

En el Hotel W ubicado en el polo turístico de Santiago de Cuba, el Departamento de Calidad decidió aplicar una encuesta a los clientes externos, para conocer cómo ha sido percibida la calidad de los servicios que ofrecen los cuatro restaurantes de la instalación. El equipo que labora en el departamento, previamente determinó el tamaño de muestra a encuestar, y luego, a la cantidad de clientes determinada, le administró el cuestionario. Este último ha sido un instrumento compuesto por 22 ítems o atributos agrupados en 5 dimensiones, más un atributo 23 o global. Cada cliente de la muestra (20 en total) emitió su criterio acerca del grado de presencia percibido de cada uno de los 22 atributos (afirmaciones) en cada restaurante, utilizando una escala de tipo Likert. El cuestionario se observa a continuación:

El siguiente grupo de declaraciones se refiere a lo que usted piensa acerca del servicio de ***restauración*** *recibido en nuestro hotel. Para cada una de*

ellas indíquenos, por favor, hasta qué punto considera que el servicio recibido posee las características descritas. Si en el lugar correspondiente usted coloca un 1, eso significa que está totalmente en desacuerdo con que el servicio recibido posee esa característica. Si coloca un 5, significa que usted está totalmente de acuerdo con la declaración. Usted puede poner un número intermedio que mejor represente sus opiniones al respecto.

1 ..... 2 ..... 3 ..... 4 5
Totalmente en desacuerdo ***Totalmente de acuerdo***

Leyenda: R *(restaurantes)*

R_1: *restaurante italiano "La Fontana"*
R_2: *restaurante de especialidades "Mariposa Blanca"*
R_3: *restaurante steak house "El Bambú"*
R_4: *restaurante de comida criolla "El Adoquín"*

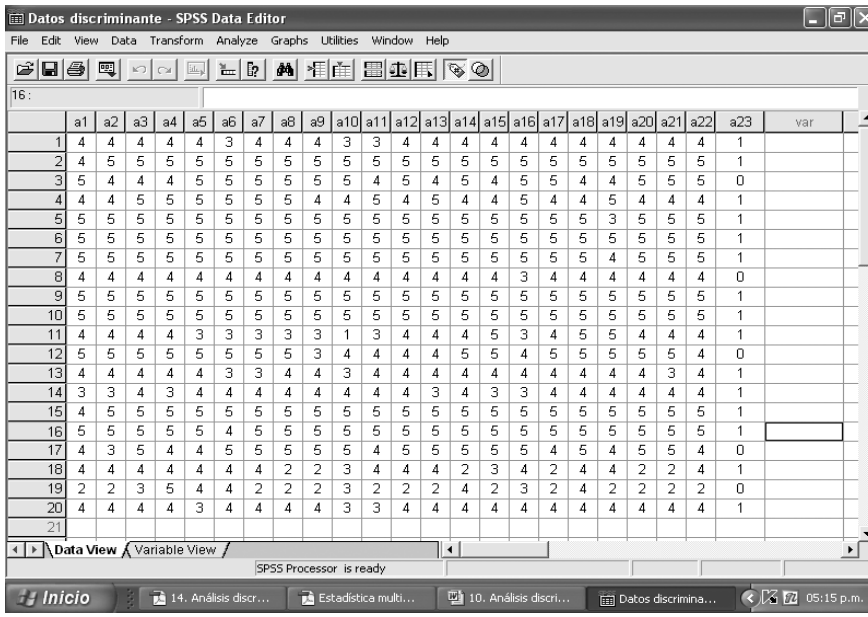
			Restaurantes			
Dimensión	**No.**	**Ítems**	**R_1**	**R_2**	**R_3**	**R_4**
1 (Elementos tangibles)	**1**	El servicio posee equipos de apariencia moderna				
	2	El restaurante es visualmente atractivo				
	3	Los empleados poseen apariencia pulcra				
	4	Los elementos materiales relacionados con el servicio son visualmente atractivos				
2 (Fiabilidad)	**5**	Cuando prometieron hacer algo en cierto tiempo, lo cumplieron				
	6	Cuando un cliente tuvo un problema, los prestadores mostraron interés en solucionarlo				
	7	Los prestadores realizaron el servicio bien a la primera				
	8	Los prestadores concluyeron el servicio en el tiempo prometido				
	9	Los prestadores del servicio no cometieron errores				
3 (Responsa-bilidad)	**10**	Los prestadores comunicaron a los clientes cuándo concluía el servicio				
	11	Los empleados ofrecieron un servicio rápido				
	12	Los empleados siempre estuvieron dispuestos a ayudar a los clientes				
	13	Los empleados siempre respondieron las preguntas de los clientes				
4 (Seguridad)	**14**	El comportamiento de los empleados me transmitió confianza				
	15	Me sentí seguro en mi relación con los empleados				
	16	Los empleados fueron siempre amables				
	17	Los empleados mostraron conocimientos suficientes para responder a las preguntas				

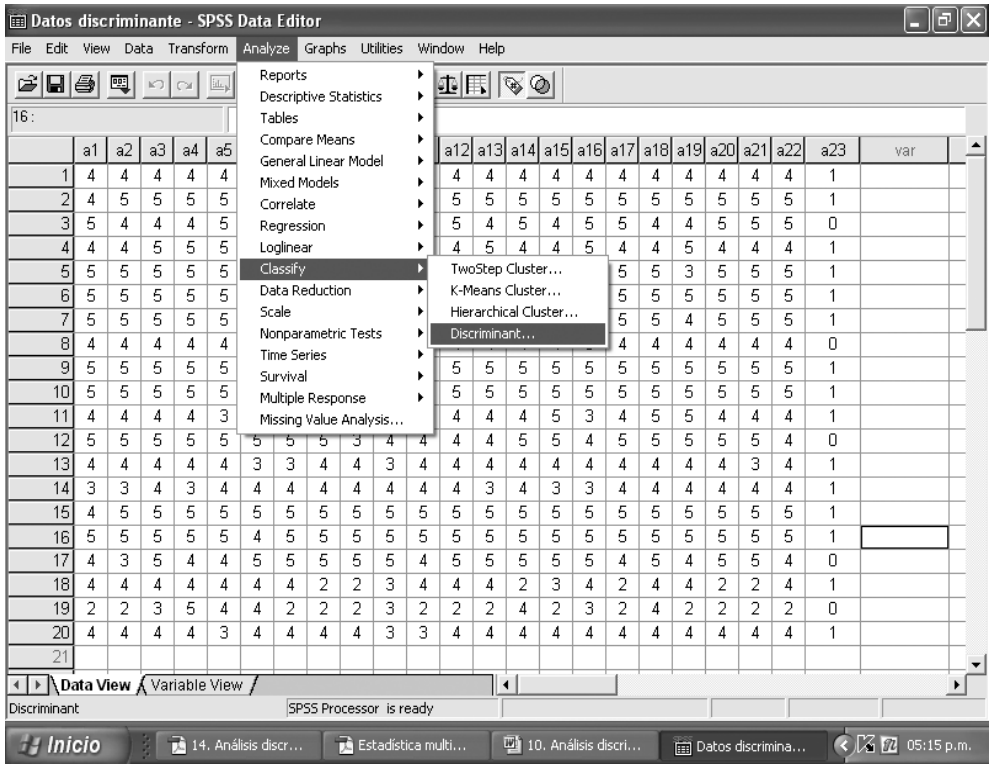
5 (Empatía)	**18**	Los empleados dieron una atención personalizada a los clientes				
	19	El servicio cuenta con un horario conveniente				
	20	El personal fue cortés en sus relaciones con los clientes				
	21	Los empleados mostraron preocupación por los intereses de los clientes				
	22	Los empleados mostraron comprensión sobre las necesidades de los clientes				
	23	**¿Recomendaría a otros nuestros servicios de restauración? Sí ____ No ____**				

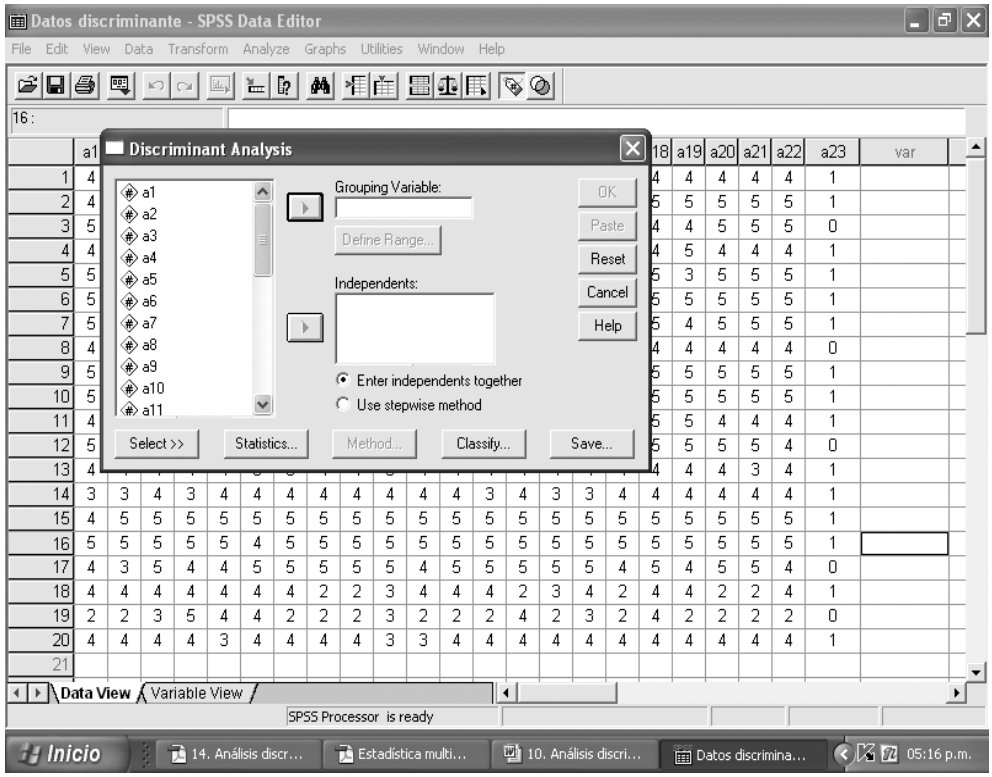
Después que el equipo del departamento aplicó la encuesta, tabuló en una base de datos las respuestas de los 20 clientes externos a cada uno de los 22 ítems. Ahora analizarán, cuáles de estos últimos, contribuyen en mayor grado a diferenciar a los clientes en los dos grupos definidos a priori (sí recomendar los servicios, no recomendar: ítem 23).

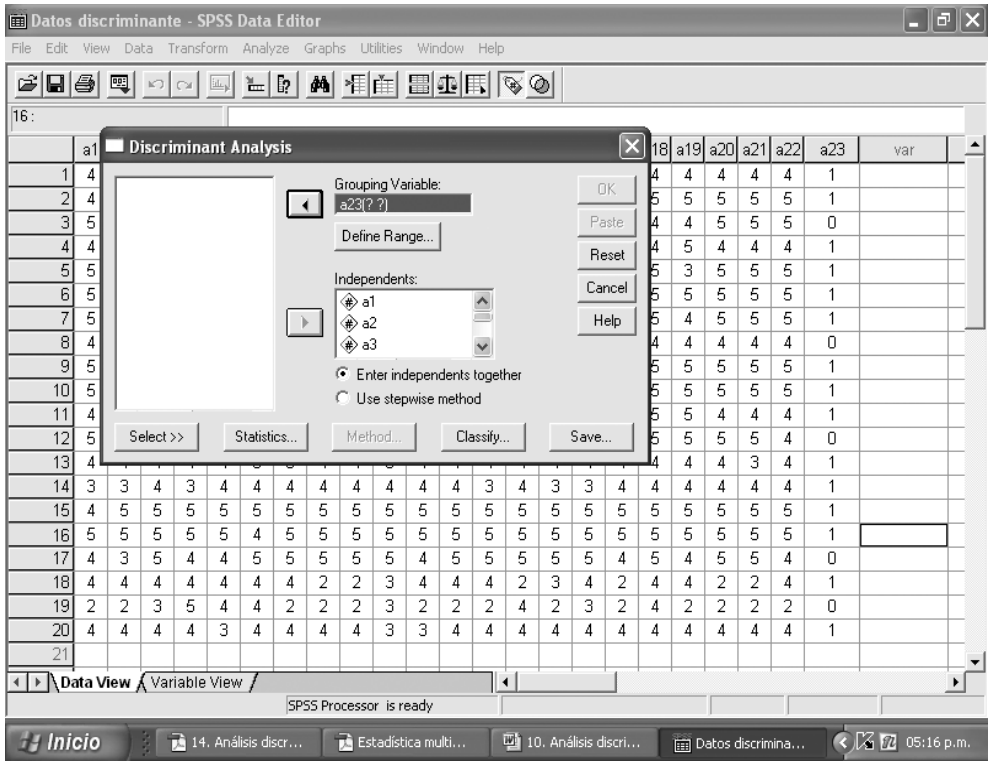
Solución:
Variables independientes: 22 atributos del cuestionario
Variable dependiente (categórica) dividida en dos grupos: sí recomendar, no recomendar
Sépase que las dos posibles respuestas (sí, no) a la variable dependiente, deben codificarse para que el software pueda procesar los datos. Se ha decidido codificar de la siguiente forma:

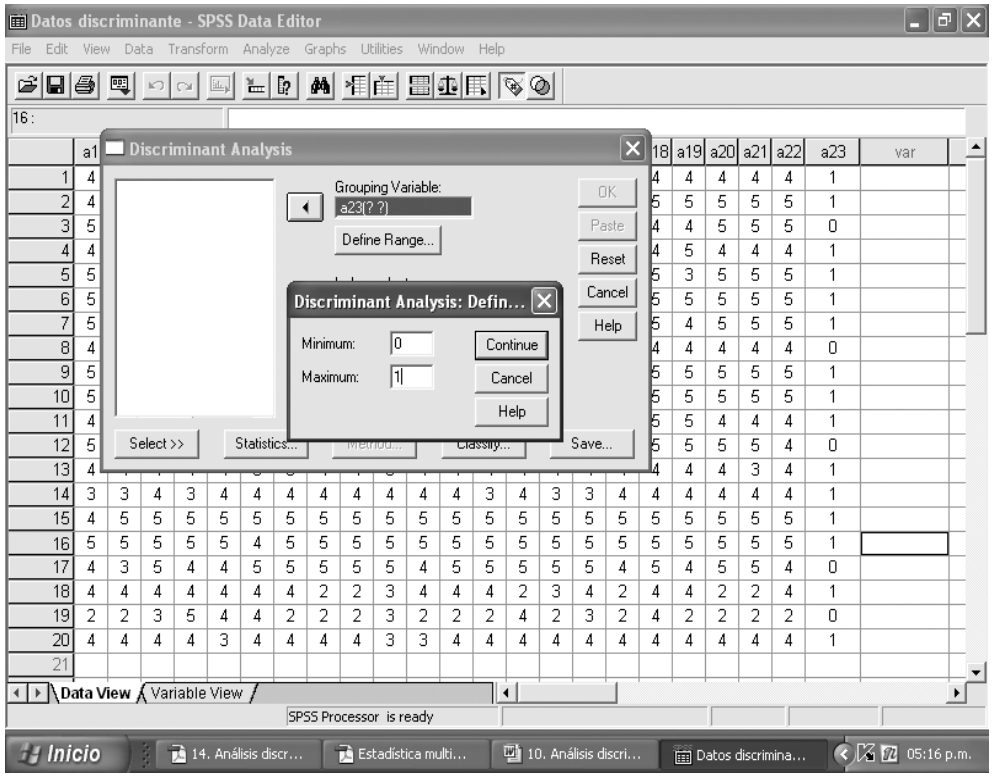
No = 0
Sí = 1
Colocando la base de datos en el SPSS, sería:

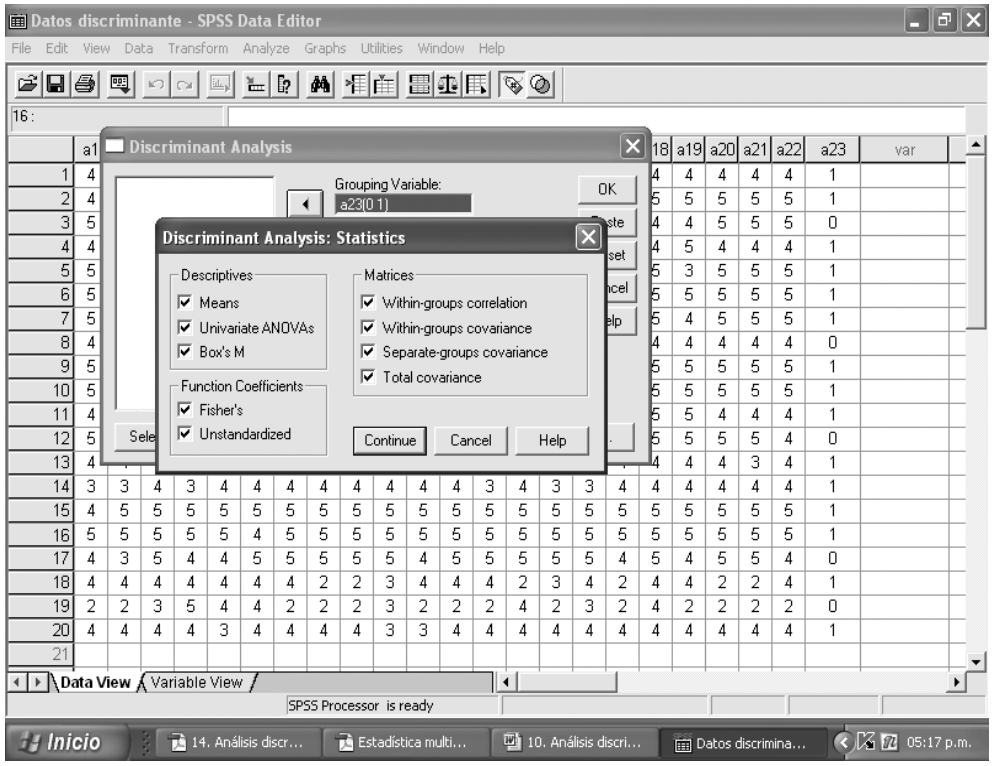


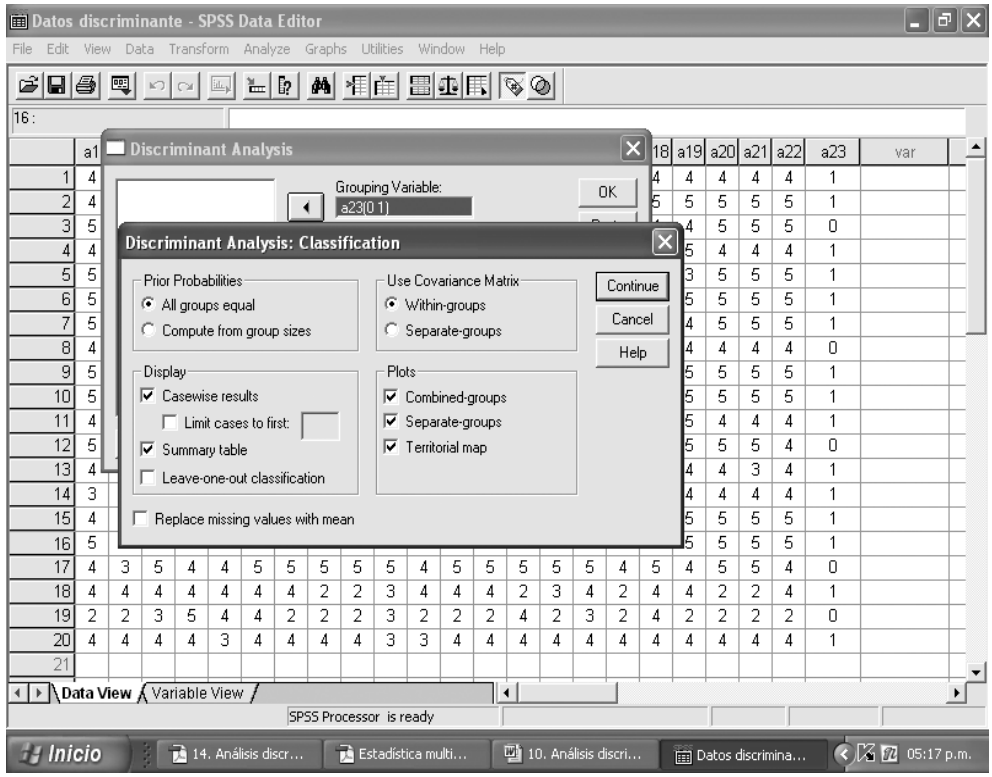

Datos discriminante - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
TwoStep Cluster...
K-Means Cluster...
Hierarchical Cluster...
Discriminant...
Data View
Variable View
Discriminant
SPSS Processor is ready
Inicio
14. Análisis discr...
Estadística multi...
10. Análisis discri...
Datos discrimina...
05:15 p.m.

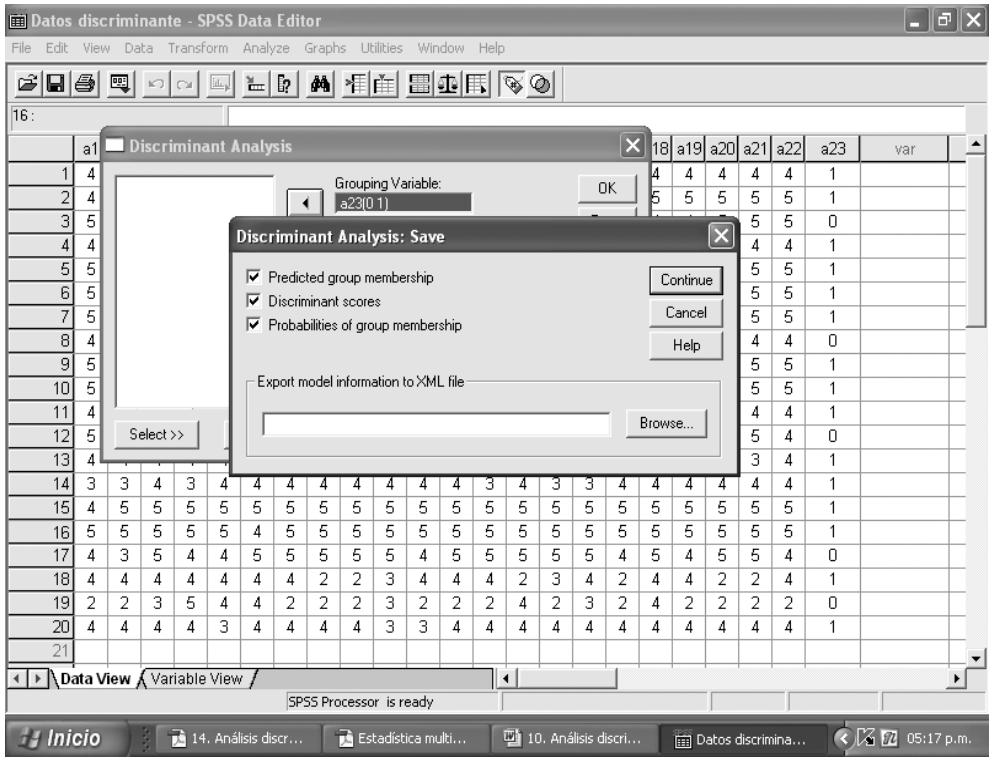

Datos discriminante - SPSS Data Editor
Discriminant Analysis
a1
a2
a3
a4
a5
a6
a7
a8
a9
a10
a11
Grouping Variable:
Define Range...
Independents:
Enter independents together
Use stepwise method
OK
Paste
Reset
Cancel
Help
Select >>
Statistics...
Method...
Classify...
Save...
Data View
Variable View
SPSS Processor is ready
Inicio
14. Análisis discr...
Estadística multi...
10. Análisis discri...
Datos discrimina...
05:16 p.m.

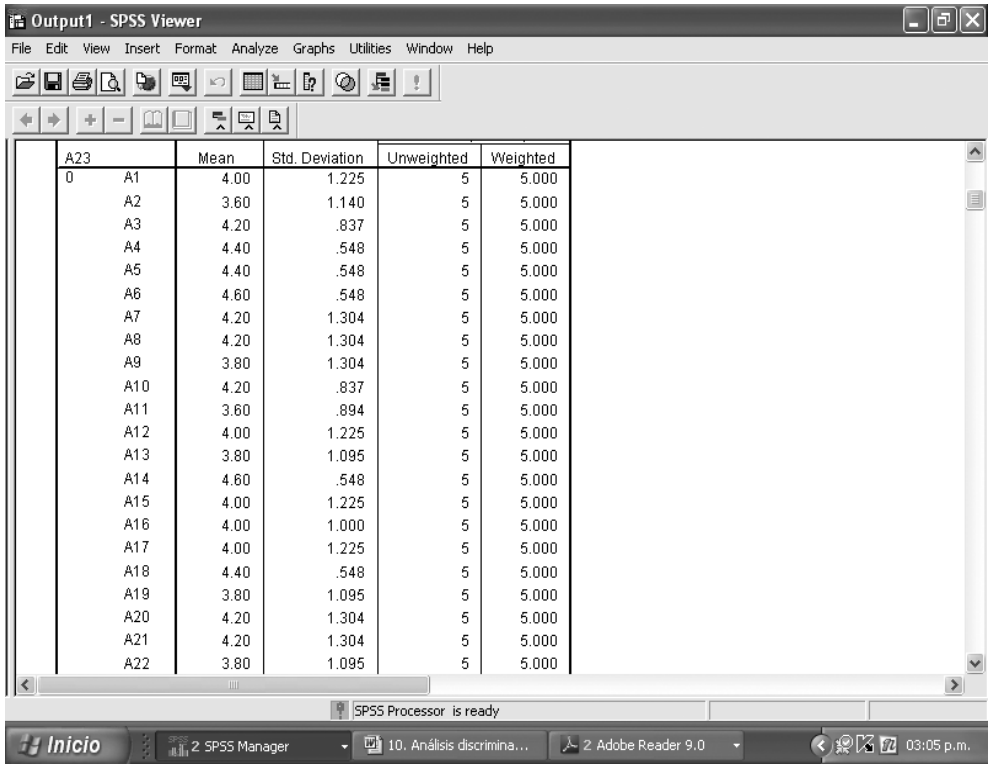

Datos discriminante - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Discriminant Analysis
Grouping Variable:
a23(? ?)
Define Range...
Independents:
a1
a2
a3
Enter independents together
Use stepwise method
OK
Paste
Reset
Cancel
Help
Select >>
Statistics...
Method...
Classify...
Save...
Data View
Variable View
SPSS Processor is ready
Inicio
14. Análisis discr...
Estadística multi...
10. Análisis discri...
Datos discrimina...
05:16 p.m.


Datos discriminante - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Discriminant Analysis
Grouping Variable:
a23(? ?)
Define Range...
Discriminant Analysis: Defin...
Minimum:
0
Maximum:
1
Continue
Cancel
Help
OK
Paste
Reset
Cancel
Help
Select >>
Statistics...
Save...
Data View
Variable View
SPSS Processor is ready
Inicio
14. Análisis discr...
Estadística multi...
10. Análisis discri...
Datos discrimina...
05:16 p.m.


Datos discriminante - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Discriminant Analysis
Grouping Variable:
a23(0 1)
OK
Discriminant Analysis: Statistics
Descriptives
Means
Univariate ANOVAs
Box's M
Function Coefficients
Fisher's
Unstandardized
Matrices
Within-groups correlation
Within-groups covariance
Separate-groups covariance
Total covariance
Continue
Cancel
Help
Data View
Variable View
SPSS Processor is ready
Inicio
14. Análisis discr...
Estadística multi...
10. Análisis discri...
Datos discrimina...
05:17 p.m.

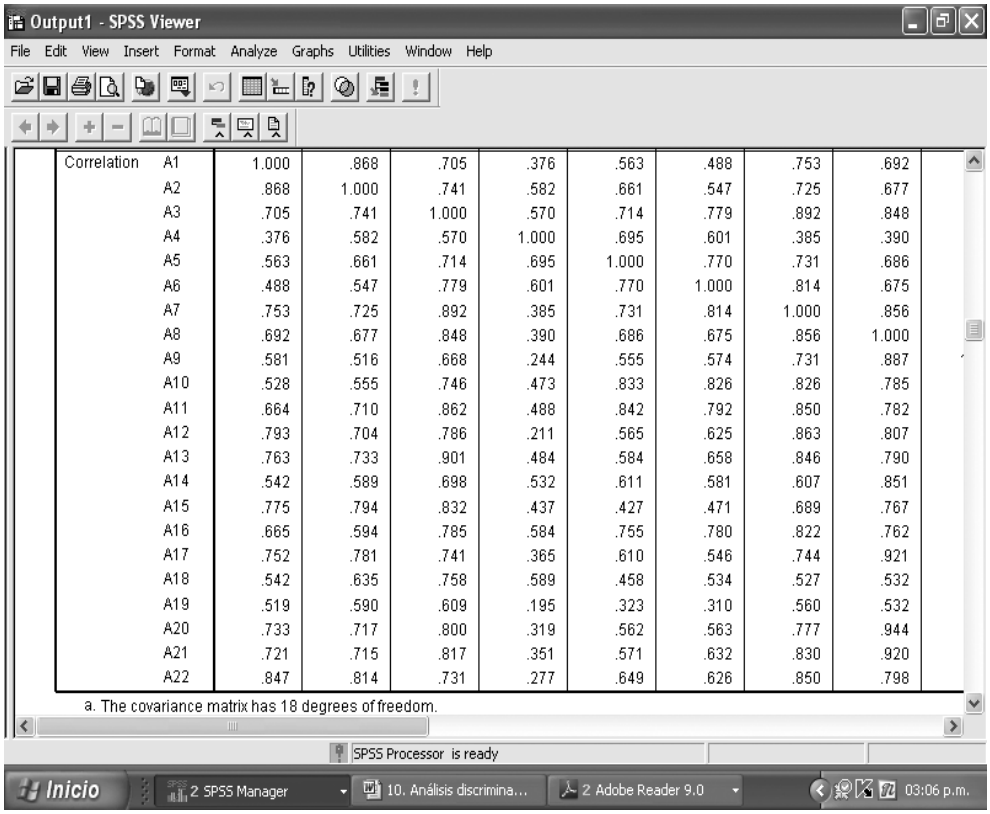

Datos discriminante - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Discriminant Analysis
Grouping Variable:
a23(0 1)
OK
Discriminant Analysis: Classification
Prior Probabilities
All groups equal
Compute from group sizes
Use Covariance Matrix
Within-groups
Separate-groups
Continue
Cancel
Help
Display
Casewise results
Limit cases to first:
Summary table
Leave-one-out classification
Plots
Combined-groups
Separate-groups
Territorial map
Replace missing values with mean
Data View
Variable View
SPSS Processor is ready
Inicio
14. Análisis discr...
Estadística multi...
10. Análisis discri...
Datos discrimina...
05:17 p.m.





A23		Mean	Std. Deviation	Unweighted	Weighted
0	A1	4.00	1.225	5	5.000
	A2	3.60	1.140	5	5.000
	A3	4.20	.837	5	5.000
	A4	4.40	.548	5	5.000
	A5	4.40	.548	5	5.000
	A6	4.60	.548	5	5.000
	A7	4.20	1.304	5	5.000
	A8	4.20	1.304	5	5.000
	A9	3.80	1.304	5	5.000
	A10	4.20	.837	5	5.000
	A11	3.60	.894	5	5.000
	A12	4.00	1.225	5	5.000
	A13	3.80	1.095	5	5.000
	A14	4.60	.548	5	5.000
	A15	4.00	1.225	5	5.000
	A16	4.00	1.000	5	5.000
	A17	4.00	1.225	5	5.000
	A18	4.40	.548	5	5.000
	A19	3.80	1.095	5	5.000
	A20	4.20	1.304	5	5.000
	A21	4.20	1.304	5	5.000
	A22	3.80	1.095	5	5.000

En la imagen anterior, aparece la tabla "*Group Statistics*" donde se observa la media y la desviación típica de los datos de cada uno de los 22 atributos medidos con la escala Likert. Podría comentarse que, por ejemplo, los atributos peor percibidos por los 20 clientes en total, son el 2 (el restaurante es visualmente atractivo) y el 11 (los empleados ofrecieron un servicio rápido) que poseen una media aritmética igual a 3.6. Este valor está un poco por encima del valor medio de la escala (3: ni en desacuerdo ni de acuerdo) y cercano a 4 (de acuerdo). Asimismo, los atributos mejor percibidos son el 6 (cuando un cliente tuvo un problema, los prestadores mostraron interés sincero en solucionarlo) y el 14 (el comportamiento de los empleados me transmitió confianza) con un valor promedio igual a 4.6, cada uno. Este valor se halla por encima de 4 y muy cercano a 5 (totalmente de acuerdo).

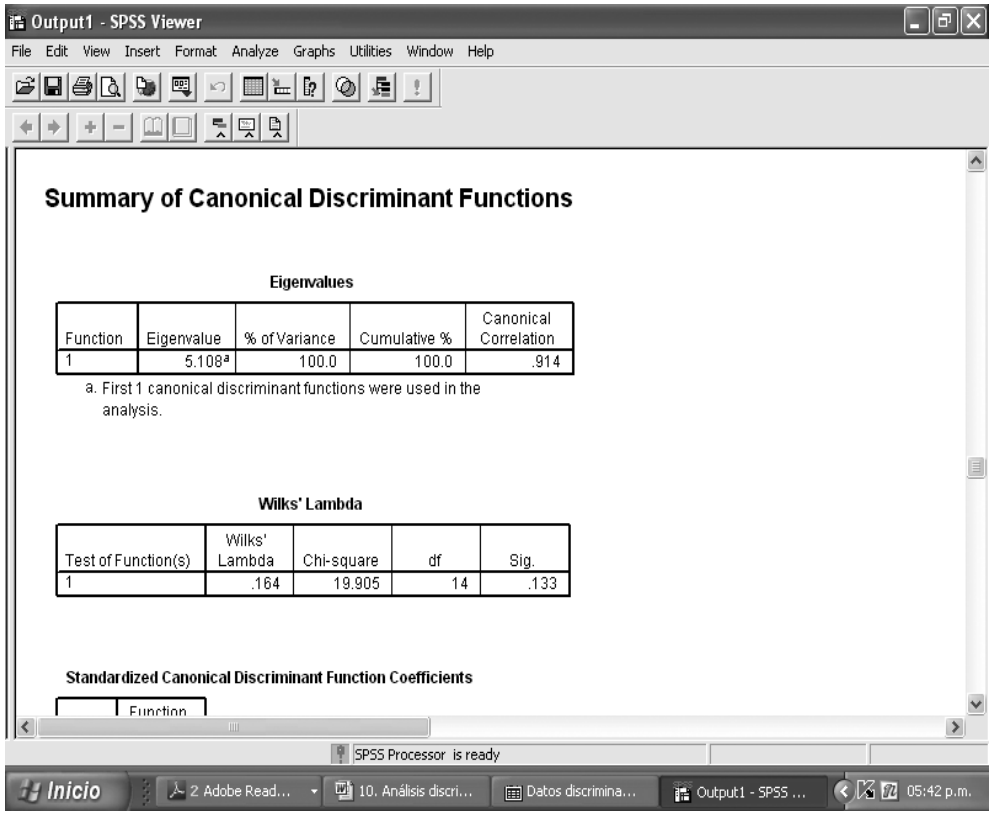


Correlation	A1	1.000	.868	.705	.376	.563	.488	.753	.692
	A2	.868	1.000	.741	.582	.661	.547	.725	.677
	A3	.705	.741	1.000	.570	.714	.779	.892	.848
	A4	.376	.582	.570	1.000	.695	.601	.385	.390
	A5	.563	.661	.714	.695	1.000	.770	.731	.686
	A6	.488	.547	.779	.601	.770	1.000	.814	.675
	A7	.753	.725	.892	.385	.731	.814	1.000	.856
	A8	.692	.677	.848	.390	.686	.675	.856	1.000
	A9	.581	.516	.668	.244	.555	.574	.731	.887
	A10	.528	.555	.746	.473	.833	.826	.826	.785
	A11	.664	.710	.862	.488	.842	.792	.850	.782
	A12	.793	.704	.786	.211	.565	.625	.863	.807
	A13	.763	.733	.901	.484	.584	.658	.846	.790
	A14	.542	.589	.698	.532	.611	.581	.607	.851
	A15	.775	.794	.832	.437	.427	.471	.689	.767
	A16	.665	.594	.785	.584	.755	.780	.822	.762
	A17	.752	.781	.741	.365	.610	.546	.744	.921
	A18	.542	.635	.758	.589	.458	.534	.527	.532
	A19	.519	.590	.609	.195	.323	.310	.560	.532
	A20	.733	.717	.800	.319	.562	.563	.777	.944
	A21	.721	.715	.817	.351	.571	.632	.830	.920
	A22	.847	.814	.731	.277	.649	.626	.850	.798

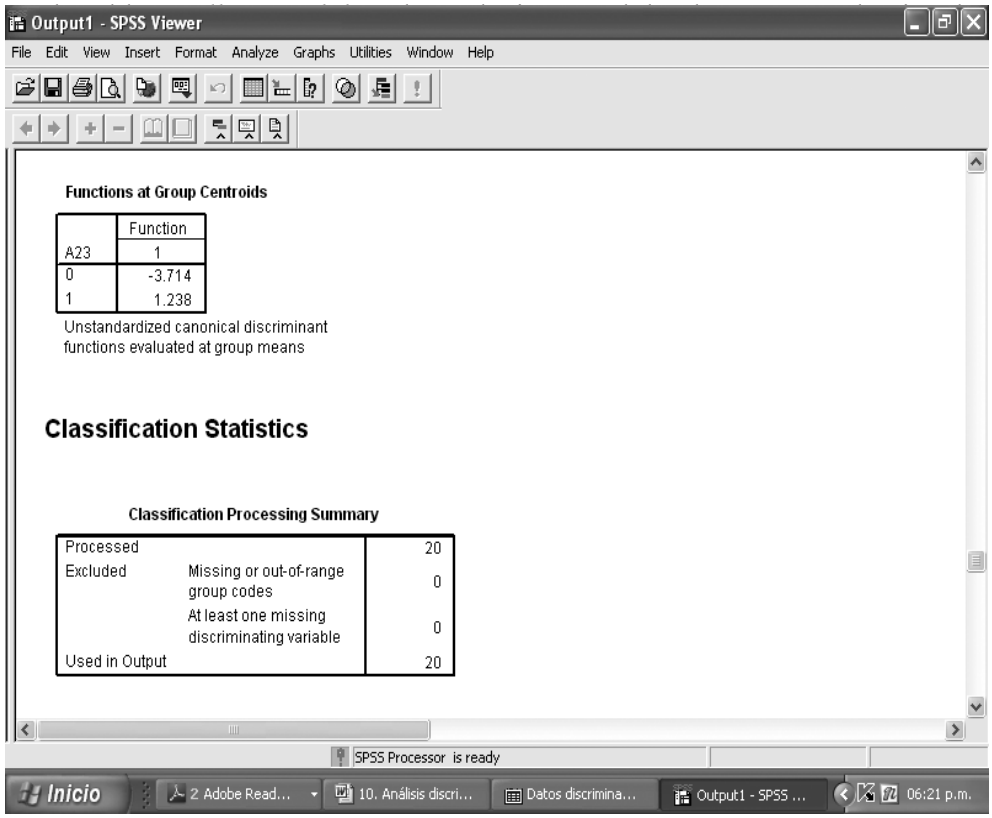
a. The covariance matrix has 18 degrees of freedom.

En la imagen anterior, aparece la tabla "*Pooled Within-Groups Matrices*" donde se observa la correlación que existe entre cada uno de los 22 atributos o variables independientes. Por ejemplo, los atributos que menos se relacionan son el 4 (los elementos materiales relacionados con el servicio son visualmente

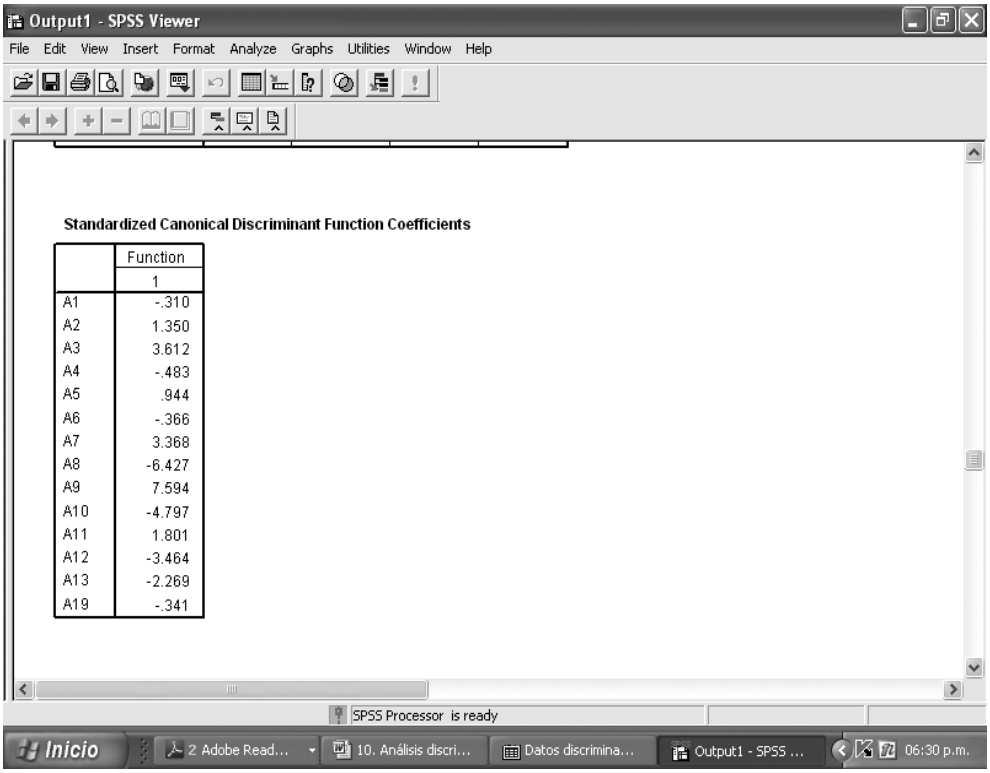
atractivos) con el 19 (el servicio cuenta con un horario conveniente) pues el coeficiente de correlación es igual a 0.195, muy cercano a 0. Por el contrario, los que se correlacionan más fuertemente son el 17 (los empleados mostraron conocimientos suficientes para responder a las preguntas) con el 20 (el personal fue cortés en sus relaciones con los clientes) pues poseen un coeficiente de correlación igual a 0.975, muy cercano a 1.



En la imagen anterior, aparece la tabla "*Eigenvalues*" donde el coeficiente de correlación canónica posee un valor muy elevado (0.914) cercano a 1, indicando que las variables discriminantes (22 atributos) permiten diferenciar entre los dos grupos (los clientes que no recomiendan y los que sí recomiendan el servicio). Igualmente, el autovalor (eigenvalue) lejano de cero (5.108) reafirma lo dicho anteriormente.



En la imagen anterior, aparece la tabla "*Functions at Group Centroids*" donde el grupo de los clientes que no recomiendan el servicio (código 0), tiende a obtener puntuaciones negativas (-3.714), mientras que el grupo de los clientes que sí recomiendan el servicio (código 1), tiende a alcanzar puntuaciones positivas (1.238). Este razonamiento sirve para comprender mejor la información que ofrece la imagen siguiente:



En la imagen anterior, aparece la tabla "*Standardized Canonical Discriminant Function Coefficients*" donde se puede afirmar que un valor por encima de la media en los atributos 2, 3, 5, 7, 9 y 11 (valores positivos), hará más probable que un cliente se ajuste al patrón de los que sí recomiendan el servicio, mientras que un valor por debajo, hará más probable que un cliente se ajuste al patrón de los que no recomiendan el servicio.

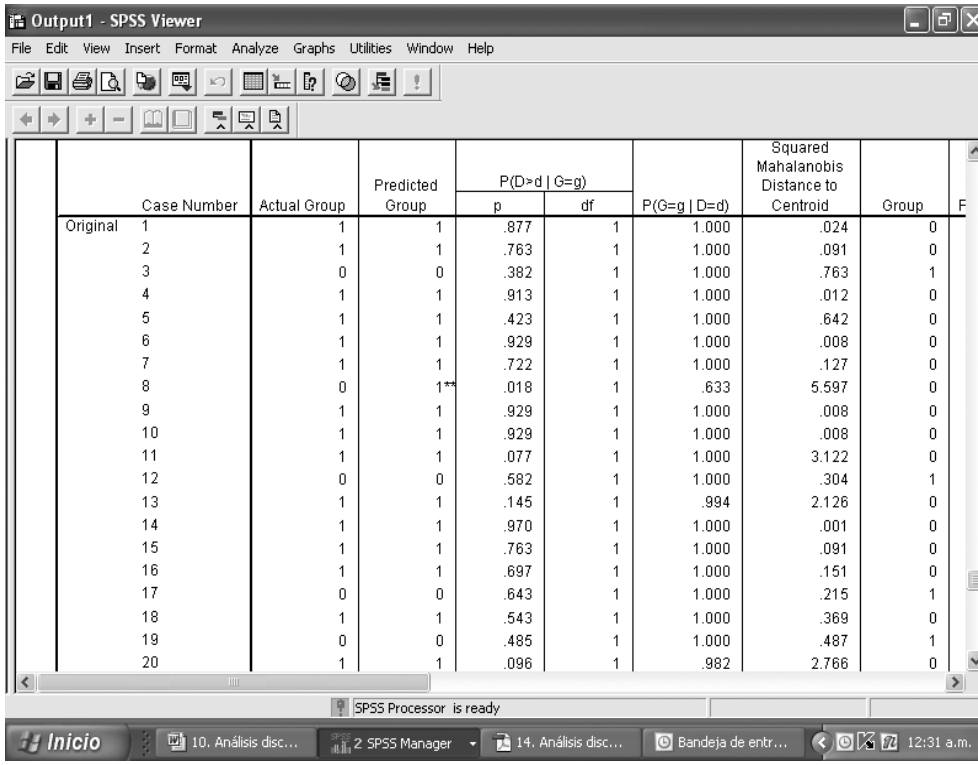
Atributo 2: el restaurante es visualmente atractivo
Atributo 3: los empleados poseen apariencia pulcra
Atributo 5: cuando prometieron hacer algo en cierto tiempo, lo cumplieron
Atributo 7: los prestadores realizaron el servicio bien a la primera
Atributo 9: los prestadores del servicio no cometieron errores
Atributo 11: los empleados ofrecieron un servicio rápido

Asimismo, un valor por encima de la media en los atributos 1, 4, 6, 8, 10, 12, 13 y 19 (valores negativos) hará más probable que un cliente se ajuste al patrón de los que no recomiendan el servicio, mientras que un valor por debajo, hará más probable que un cliente se ajuste al patrón de los que sí recomiendan el servicio.

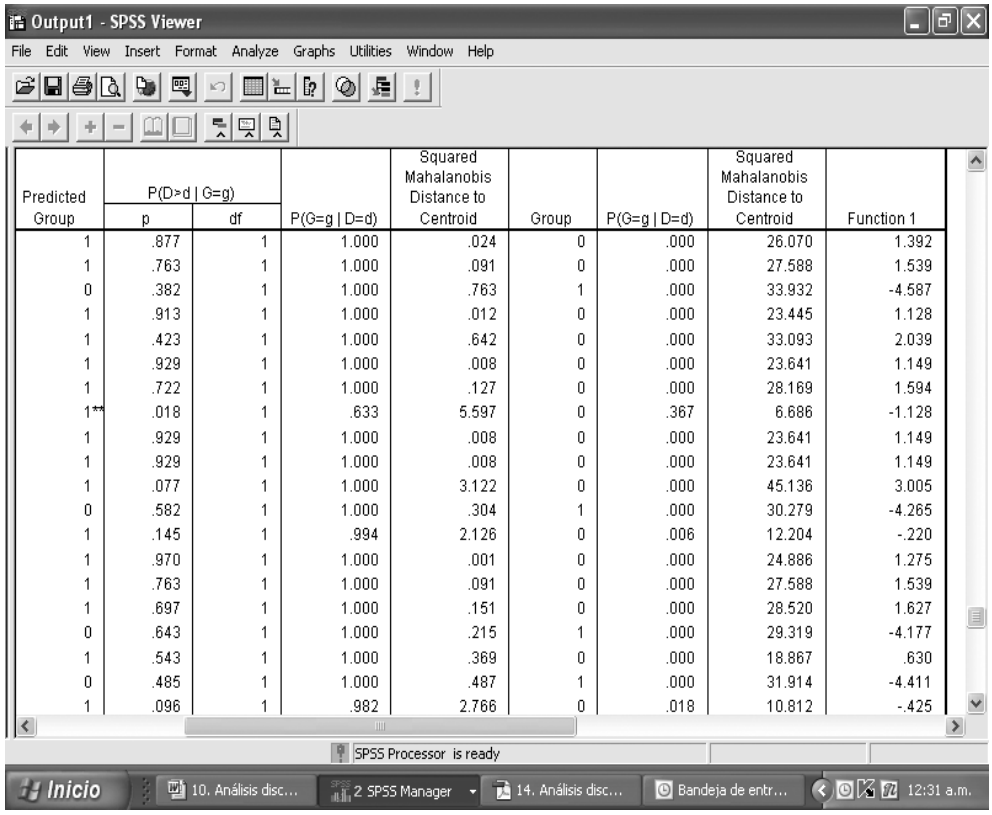
Atributo 1: el servicio posee equipos de apariencia moderna
Atributo 4: los elementos materiales relacionados con el servicio, son visualmente atractivos
Atributo 6: cuando un cliente tuvo un problema, los prestadores mostraron interés sincero en solucionarlo
Atributo 8: los prestadores concluyeron el servicio en el tiempo prometido
Atributo 10: los prestadores comunicaron a los clientes cuándo concluía el servicio
Atributo 12: los empleados siempre estuvieron dispuestos a ayudar a los clientes
Atributo 13: los empleados siempre respondieron las preguntas de los clientes
Atributo 19: el servicio cuenta con un horario conveniente

Nótese que se ha obtenido una sola función discriminante (en la tabla anterior, representada como *function 1*), pues la cantidad de funciones responde a K-1 grupos. Como son dos grupos definidos a priori (sí recomendar, no recomendar) pues se obtendrá una sola función.

Nótese también que como parte del análisis discriminante, se han escogido los atributos del cuestionario que permiten diferenciar a los grupos, pues sólo éstos son los necesarios para alcanzar la mejor clasificación posible. Ellos han sido desde el atributo 1 hasta el 13 y además el 19.

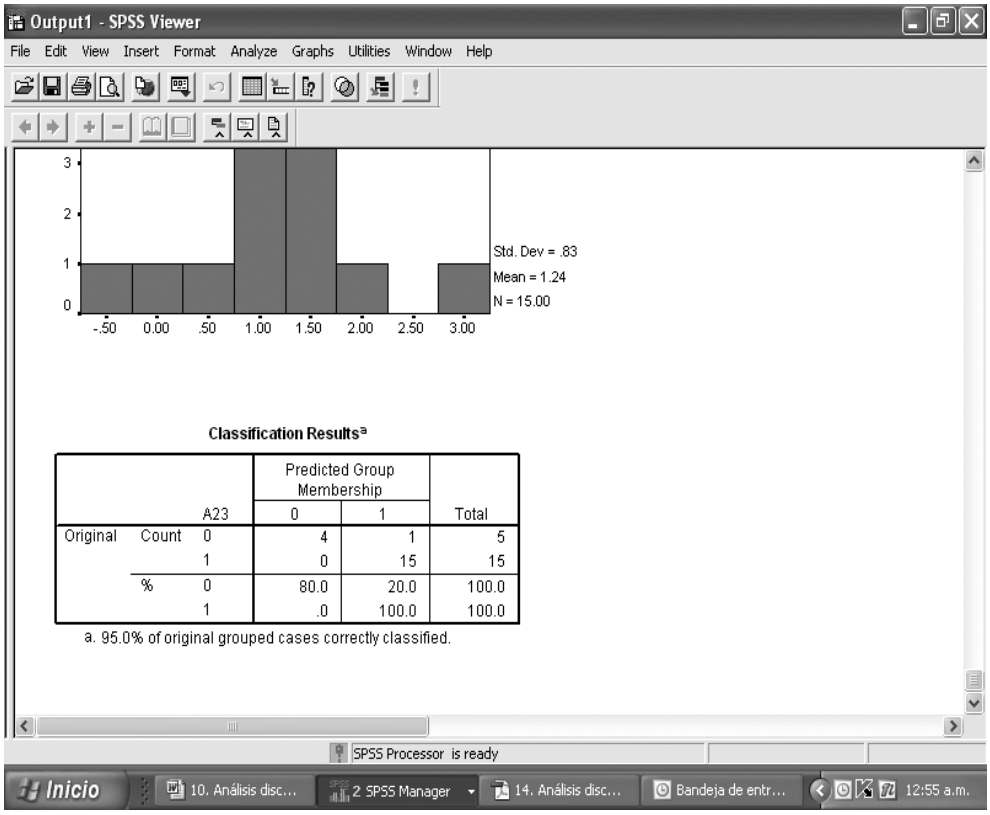


	Case Number	Actual Group	Predicted Group	P(D>d | G=g) p	P(D>d | G=g) df	P(G=g | D=d)	Squared Mahalanobis Distance to Centroid	Group	F
Original	1	1	1	.877	1	1.000	.024	0	
	2	1	1	.763	1	1.000	.091	0	
	3	0	0	.382	1	1.000	.763	1	
	4	1	1	.913	1	1.000	.012	0	
	5	1	1	.423	1	1.000	.642	0	
	6	1	1	.929	1	1.000	.008	0	
	7	1	1	.722	1	1.000	.127	0	
	8	0	1**	.018	1	.633	5.597	0	
	9	1	1	.929	1	1.000	.008	0	
	10	1	1	.929	1	1.000	.008	0	
	11	1	1	.077	1	1.000	3.122	0	
	12	0	0	.582	1	1.000	.304	1	
	13	1	1	.145	1	.994	2.126	0	
	14	1	1	.970	1	1.000	.001	0	
	15	1	1	.763	1	1.000	.091	0	
	16	1	1	.697	1	1.000	.151	0	
	17	0	0	.643	1	1.000	.215	1	
	18	1	1	.543	1	1.000	.369	0	
	19	0	0	.485	1	1.000	.487	1	
	20	1	1	.096	1	.982	2.766	0	

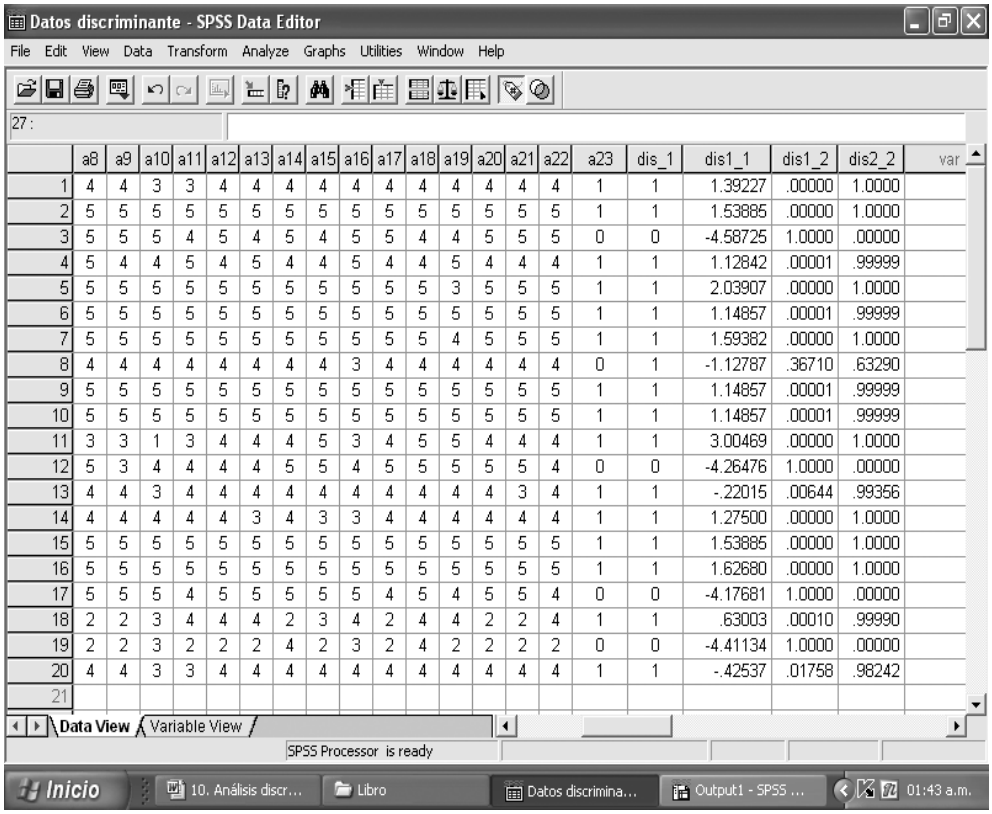


Predicted Group	P(D>d | G=g) p	P(D>d | G=g) df	P(G=g | D=d)	Squared Mahalanobis Distance to Centroid	Group	P(G=g | D=d)	Squared Mahalanobis Distance to Centroid	Function 1
1	.877	1	1.000	.024	0	.000	26.070	1.392
1	.763	1	1.000	.091	0	.000	27.588	1.539
0	.382	1	1.000	.763	1	.000	33.932	-4.587
1	.913	1	1.000	.012	0	.000	23.445	1.128
1	.423	1	1.000	.642	0	.000	33.093	2.039
1	.929	1	1.000	.008	0	.000	23.641	1.149
1	.722	1	1.000	.127	0	.000	28.169	1.594
1**	.018	1	.633	5.597	0	.367	6.686	-1.128
1	.929	1	1.000	.008	0	.000	23.641	1.149
1	.929	1	1.000	.008	0	.000	23.641	1.149
1	.077	1	1.000	3.122	0	.000	45.136	3.005
0	.582	1	1.000	.304	1	.000	30.279	-4.265
1	.145	1	.994	2.126	0	.006	12.204	-.220
1	.970	1	1.000	.001	0	.000	24.886	1.275
1	.763	1	1.000	.091	0	.000	27.588	1.539
1	.697	1	1.000	.151	0	.000	28.520	1.627
0	.643	1	1.000	.215	1	.000	29.319	-4.177
1	.543	1	1.000	.369	0	.000	18.867	.630
0	.485	1	1.000	.487	1	.000	31.914	-4.411
1	.096	1	.982	2.766	0	.018	10.812	-.425

En las dos imágenes anteriores, aparece la tabla "*Casewise Statistics*" donde el cliente número 8 pertenece al grupo 0 (no recomendar) pero ha sido clasificado en el grupo 1 (sí recomendar). Esto se debe a que la puntuación discriminante para este sujeto de la muestra, equivale a -1.128 el cual es un número más cercano al grupo 1 (sí recomendar = 1.238) que al grupo 0 (no recomendar = -3.714).



En la imagen anterior, aparece la tabla "*Classification Results*" donde se muestra que los clientes que no recomiendan el servicio (4 en total), son correctamente clasificados en un 80%, y que los clientes que sí recomiendan el servicio (15 en total), lo son en un 100%. En general, la función discriminante consigue clasificar correctamente al 95% de los clientes encuestados, un por ciento bastante elevado que permite, además, comprobar que la función discriminante obtenida, posee un alto grado de eficacia desde el punto de vista de la clasificación. Obsérvese que el 5% restante, representa un solo encuestado (1 de 20), el número 8 que originalmente respondió erróneamente al decir que no recomienda el servicio.



	a8	a9	a10	a11	a12	a13	a14	a15	a16	a17	a18	a19	a20	a21	a22	a23	dis_1	dis1_1	dis1_2	dis2_2	var
1	4	4	3	3	4	4	4	4	4	4	4	4	4	4	4	1	1	1.39227	.00000	1.0000	
2	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	1	1	1.53885	.00000	1.0000	
3	5	5	5	4	5	4	5	4	5	5	4	4	5	5	5	0	0	-4.58725	1.0000	.00000	
4	5	4	4	5	4	5	4	4	5	4	4	5	4	4	4	1	1	1.12842	.00001	.99999	
5	5	5	5	5	5	5	5	5	5	5	5	3	5	5	5	1	1	2.03907	.00000	1.0000	
6	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	1	1	1.14857	.00001	.99999	
7	5	5	5	5	5	5	5	5	5	5	5	4	5	5	5	1	1	1.59382	.00000	1.0000	
8	4	4	4	4	4	4	4	4	3	4	4	4	4	4	4	0	1	-1.12787	.36710	.63290	
9	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	1	1	1.14857	.00001	.99999	
10	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	1	1	1.14857	.00001	.99999	
11	3	3	1	3	4	4	4	5	3	4	5	5	4	4	4	1	1	3.00469	.00000	1.0000	
12	5	3	4	4	4	4	5	5	4	5	5	5	5	5	4	0	0	-4.26476	1.0000	.00000	
13	4	4	3	4	4	4	4	4	4	4	4	4	4	3	4	1	1	-.22015	.00644	.99356	
14	4	4	4	4	4	3	4	3	3	4	4	4	4	4	4	1	1	1.27500	.00000	1.0000	
15	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	1	1	1.53885	.00000	1.0000	
16	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	1	1	1.62680	.00000	1.0000	
17	5	5	5	4	5	5	5	5	5	4	5	4	5	5	4	0	0	-4.17681	1.0000	.00000	
18	2	2	3	4	4	4	2	3	4	2	4	4	2	2	4	1	1	.63003	.00010	.99990	
19	2	2	3	2	2	2	4	2	3	2	4	2	2	2	2	0	0	-4.41134	1.0000	.00000	
20	4	4	3	3	4	4	4	4	4	4	4	4	4	4	4	1	1	-.42537	.01758	.98242	
21																					

Véase, por último, la matriz de datos original colocada en el SPSS tal y como muestra la imagen anterior. El propio software ha creado cuatro nuevas columnas o variables con datos que ofrecen una información importante para entender la clasificación de los clientes en cada grupo. Estas se denominan:

- dis_1
- dis1_1
- dis1_2
- dis2_2

La variable dis_1 representa el grupo asignado a cada cliente.
La variable dis1_1 representa la puntuación discriminante para cada cliente.
La variable dis1_2 representa la probabilidad que tiene cada cliente de pertenecer al grupo uno (no recomendar).
La variable dis2_2 representa la probabilidad que tiene cada cliente de pertenecer al grupo dos (sí recomendar).
Esto es muy importante, pues al observar los datos de las cuatro columnas,

se comprende la información que ofrece la tabla última de resultados "*Classification Results*". Véase, por ejemplo, que de los veinte clientes clasificados originalmente, sólo uno (el número 8) fue incorrectamente clasificado. Esto se aprecia en la base de datos al observar que ese cliente, dijo que no recomendaba (0) el servicio (columna: a23), cuando según su respuesta a los 22 atributos del cuestionario mostrando alta percepción de calidad, debió decir que sí recomendaba (1) el servicio (columna: dis_1). El valor de la puntuación discriminante igual a -1.12787 hace ver que el mismo está más cercano a 1.238 (sí recomendar) que a -3.714 (no recomendar). Es por ello que la probabilidad de pertenecer al grupo dos (sí recomendar) igual a 0.63290 es mayor que la probabilidad de ser clasificado en el grupo uno (no recomendar) igual a 0.36710.

A partir de ahora y sobre la base de esas puntuaciones discriminantes como parte de la función discriminante, será posible clasificar nuevos clientes que sean encuestados, lo mismo si dan respuesta (correcta o errónea) de recomendar o no el servicio, como si no dan respuesta alguna en ese atributo 23 o variable dependiente.

EJERCITACIÓN

En cierto polo turístico, la Delegación del Turismo ha recopilado los datos correspondientes a 19 variables medidas en 15 hoteles del territorio. Dichas variables son las siguientes:

- *Variable 1:* cantidad de habitaciones
- *Variable 2:* precio por habitación
- *Variable 3:* cantidad de restaurantes especializados
- *Variable 4:* cantidad de bares
- *Variable 5:* cantidad de empleados
- *Variable 6:* nivel de lujo (1: bajo, 2: medio, 3: alto)
- *Variable 7:* promedio de estancia de los clientes (días)
- *Variable 8:* % de ocupación promedio
- *Variable 9:* cantidad de servicios de animación
- *Variable 10:* nivel de calidad de las comidas y bebidas (1: bajo, 2: medio, 3: alto)
- *Variable 11:* categoría de la zona de playa (1: mala, 2: regular, 3: buena)
- *Variable 12:* gasto diario promedio por cliente
- *Variable 13:* ingreso diario promedio del hotel
- *Variable 14:* tiempo de demora del servicio de check-in (minutos)
- *Variable 15:* tiempo de demora del servicio de check-out (minutos)
- *Variable 16:* nivel de profesionalidad de los empleados (1: bajo, 2: medio, 3: alto)
- *Variable 17:* cantidad de piscinas
- *Variable 18:* cantidad de servicios extra
- *Variable 19:* cantidad de idiomas que dominan los empleados

Los miembros de la Delegación, conocen que los 15 hoteles pertenecen a la Cadena Mundo Azul algunos, y otros, a la Cadena Buen Viaje. Han decidido ambos grupos de hoteles, codificarlos de la siguiente forma:

Hoteles de la Cadena Mundo Azul = 1
Hoteles de la Cadena Buen Viaje = 2

La base de datos se muestra a continuación:

Hotel	ch	pph	cre	cb	ce	nl	pec	% op	csa	nccb	czp	gdpc	idph	tdsci	tdsco	npe	cp	cse	cide	Grupos
1	392	112.95	4	3	310	2	5	58	6	2	1	56.80	4015.45	10	11	2	2	4	2	Buen Viaje
2	415	167.50	5	4	350	2	5	60	7	2	2	63.45	5432.25	9	10	3	3	5	3	Buen Viaje
3	298	99.55	4	3	230	2	3	59	6	2	1	58.90	4657.75	9	9	2	2	4	2	Buen Viaje
4	514	268.90	6	5	478	3	6	64	8	3	3	72.15	6354.90	8	9	3	4	6	4	Mundo Azul
5	678	310.15	7	6	605	3	7	66	9	3	3	81.50	7154.80	6	8	3	5	7	5	Mundo Azul
6	356	145.65	4	3	302	2	3	55	6	2	1	55.35	4657.55	10	10	2	2	4	2	Buen Viaje
7	425	171.45	5	4	397	2	3	57	7	2	2	62.40	5367.80	9	10	3	3	5	3	Buen Viaje
8	546	254.00	6	5	499	3	5	61	8	3	3	70.90	6154.50	8	9	3	4	6	4	Mundo Azul
9	256	98.60	4	3	203	2	3	57	6	2	1	54.55	4256.50	9	10	2	2	4	2	Buen Viaje
10	614	356.35	7	6	584	3	6	64	9	3	3	80.90	7154.80	5	7	3	5	7	5	Mundo Azul
11	579	250.90	6	5	523	3	6	63	8	3	3	71.25	6352.35	8	8	3	4	6	4	Mundo Azul
12	712	405.80	8	7	681	3	7	75	10	3	3	92.85	8102.60	4	5	3	6	8	6	Mundo Azul
13	456	190.25	5	4	410	2	5	55	7	3	2	60.35	5143.75	9	9	3	3	5	3	Buen Viaje
14	299	96.35	3	2	224	1	3	48	5	1	1	49.25	3628.60	11	12	1	1	3	1	Buen Viaje
15	368	133.75	4	3	312	2	3	56	6	2	1	54.60	4168.45	10	11	2	2	4	2	Buen Viaje

Ahora averiguarán, en qué se diferencian o distinguen los hoteles de la Cadena Mundo Azul, y los de la Cadena Buen Viaje.

SOLUCIÓN

- el coeficiente de correlación canónica posee un valor muy elevado (0.994) cercano a 1, indicando que las variables discriminantes (19) permiten diferenciar entre los dos grupos (hoteles de la Cadena Mundo Azul y hoteles de la Cadena Buen Viaje)
- el autovalor muy lejano de cero (81.76) reafirma lo anterior
- el estadístico Lambda alcanza un valor igual a 0.012 muy próximo a cero, indicando que existe una gran diferencia entre los dos grupos (hoteles de la Cadena Mundo Azul y hoteles de la Cadena Buen Viaje)
- el grupo de los hoteles de la Cadena Mundo Azul (código 1), tiende a obtener puntuaciones negativas (-10.31), mientras que el grupo de los hoteles de la Cadena Buen Viaje (código 2), tiende a alcanzar puntuaciones positivas (6.87)
- como parte del análisis discriminante, se han escogido las variables que permiten diferenciar a los grupos, pues sólo éstas son las necesarias para alcanzar la mejor clasificación posible. Ellas han sido: desde la 1 hasta 3, desde la 5 hasta la 8 y desde la 10 hasta la 15
- se puede afirmar que un valor por encima de la media en las variables 3, 5, 7, 10, 13, 14 y 15 (valores positivos), hará más probable que un hotel se ajuste al patrón de los de la Cadena Buen Viaje, mientras que un valor por debajo, hará más probable que un hotel se ajuste al patrón de los hoteles de la Cadena Mundo Azul
- un valor por encima de la media en los atributos 1, 2, 6, 8, 11 y 12 (valores negativos) hará más probable que un hotel se ajuste al patrón de los de la Cadena Mundo Azul, mientras que un valor por debajo, hará más probable que un hotel se ajuste al patrón de los hoteles de la Cadena Buen Viaje
- finalmente, los hoteles de la Cadena Mundo Azul (6 en total) son correctamente clasificados en un 100%, y los hoteles de la Cadena Buen Viaje (9 en total) lo son en un 100% igualmente. En general, la función discriminante consigue clasificar correctamente al 100% de los hoteles tomados en el estudio, lo cual permite comprobar que la función discriminante obtenida, posee un alto grado de eficacia desde el punto de vista de la clasificación.

Capítulo
10

Análisis cluster.

10.1. Concepto de análisis cluster.

El análisis cluster (llámese también análisis de conglomerados) es una técnica multivariante que utiliza la información de una serie de variables para cada sujeto u objeto y, conforme a estas variables, se mide la similitud entre ellos. Una vez medida la similitud, se agrupan en: grupos homogéneos internamente y diferentes entre sí.

La idea conceptual básica de este tipo de análisis, parte de suponer que en muchas ocasiones, un solo individuo u objeto, constituye una unidad de observación demasiado reducida. Se trata entonces de agrupar a los sujetos originales (u objetos) en grupos, centrando el análisis en esos grupos y no en cada uno de los individuos (u objetos).

Debe aclararse que los resultados logrados para una muestra, sólo sirven para ese diseño (su valor atañe sólo a los objetivos del investigador). Se habla de resultados en cuanto a: la elección de individuos (u objetos), variables relevantes utilizadas, criterio de similitud empleado, nivel de agrupación final elegido, etc. Existen dos tipos de análisis cluster:

- análisis cluster jerárquico
- análisis cluster K-medias

Por último, resulta útil destacar que el análisis cluster y el análisis discriminante aunque parecen muy similares, realmente no lo son. El análisis discriminante

intenta explicar una estructura, y el análisis cluster pretende determinarla.

10.2. Concepto de análisis cluster jerárquico.

El análisis cluster jerárquico, permite aglomerar tanto casos como variables, y elegir entre una gran variedad de métodos de aglomeración y medidas de distancia. En éste se procede de forma jerárquica. Es una técnica aglomerativa que comienza partiendo de los elementos muestrales individualmente considerados, y va creando grupos hasta llegar a la formación de un único grupo o conglomerado, constituido por todos los elementos de la muestra.

10.3. Concepto de análisis cluster K-medias.

El análisis cluster K-medias, es un método de agrupación de casos que se basa en las distancias existentes entre ellos en un conjunto de variables. Permite procesar un número ilimitado de casos pero utilizando un único método de aglomeración. Requiere, además, que se proponga previamente el número de conglomerados que se desea obtener.

Para muestras grandes, este método resulta más aconsejable que el jerárquico.

Es importante señalar, que esta técnica de aglomeración no permite agrupar variables a diferencia del jerárquico.

10.4. Algunas puntualizaciones de interés acerca del análisis cluster.

En la matriz de coeficientes de distancia euclídea al cuadrado (o cualquier otro tipo de medida de distancia seleccionada), los coeficientes más elevados responden a mayores distancias o mayor diferencia entre los casos analizados. Por el contrario, coeficientes con más bajo valor, corresponden a menores distancias o mayor parecido entre dichos casos.

La lectura del gráfico de carámbanos vertical, se realiza de abajo hacia arriba, de modo que la última fila, corresponde al primer nivel de agrupación de los casos, y la primera fila, al último nivel. Siempre en el último nivel, quedan

agrupados todos los casos de la muestra en un solo cluster.

Un cluster puede formarse a partir de dos casos en uno solo, o añadiendo un caso a un multicluster ya existente, o uniendo dos multicluster ya existentes.

En la tabla de aglomeraciones previstas, el valor del coeficiente a cada nivel, ayuda a decidir cuántos clusters pueden constituir la mejor solución para representar los datos.

La lectura del dendograma se realiza de izquierda a derecha donde las líneas verticales representan la unión de dos clusters. La posición de la línea vertical sobre la escala de valores de 0 a 25, indica a qué distancia los clusters se han unido.

Véase un ejemplo de análisis cluster jerárquico.

Ejemplo 1:

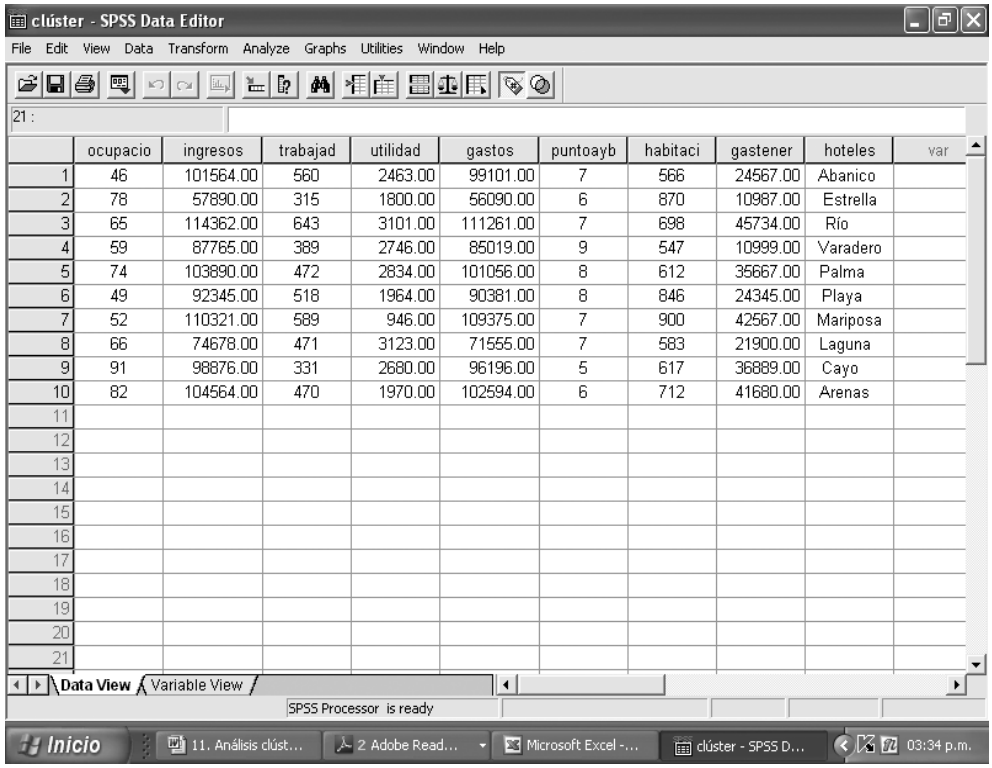
En el polo turístico de Varadero, un grupo de analistas de la Delegación del MINTUR, está realizando un estudio que incluye diez instalaciones hoteleras. Basándose en los datos recopilados de ocho variables que han sido medidas en cada uno de los diez hoteles, el objetivo de los miembros del grupo, consiste en agrupar dichas entidades según su similitud o semejanza. Los datos se muestran a continuación:

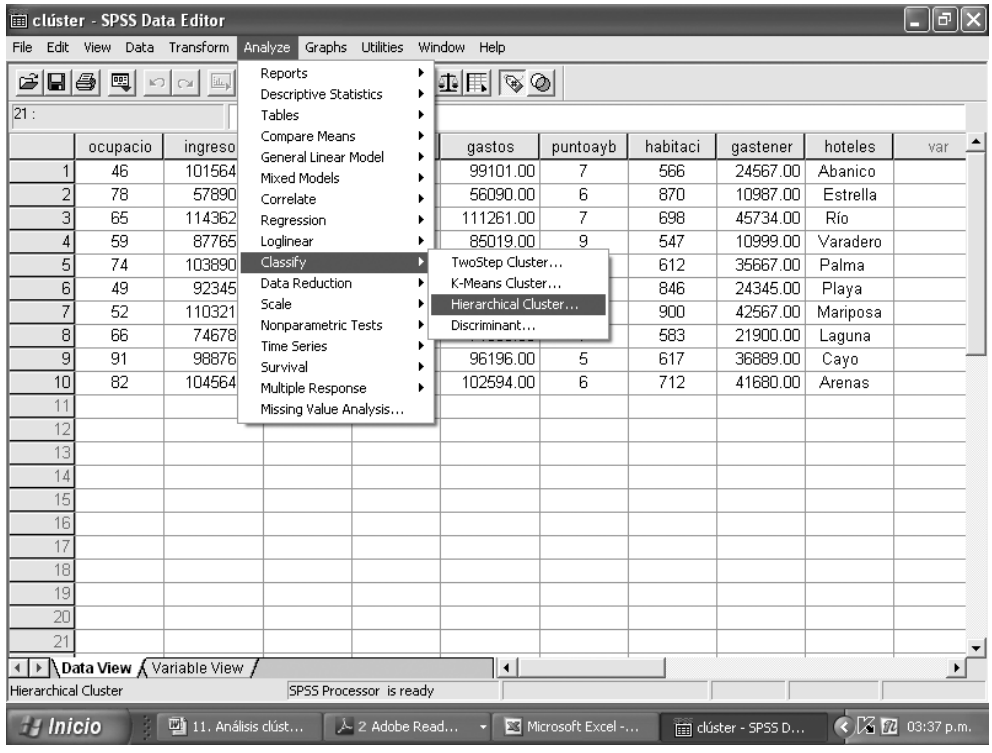
Variables:
- % de ocupación
- nivel de ingresos
- cantidad de trabajadores
- nivel de utilidades
- nivel de gastos
- cantidad de puntos de consumo de A+B
- cantidad de habitaciones
- gasto energético

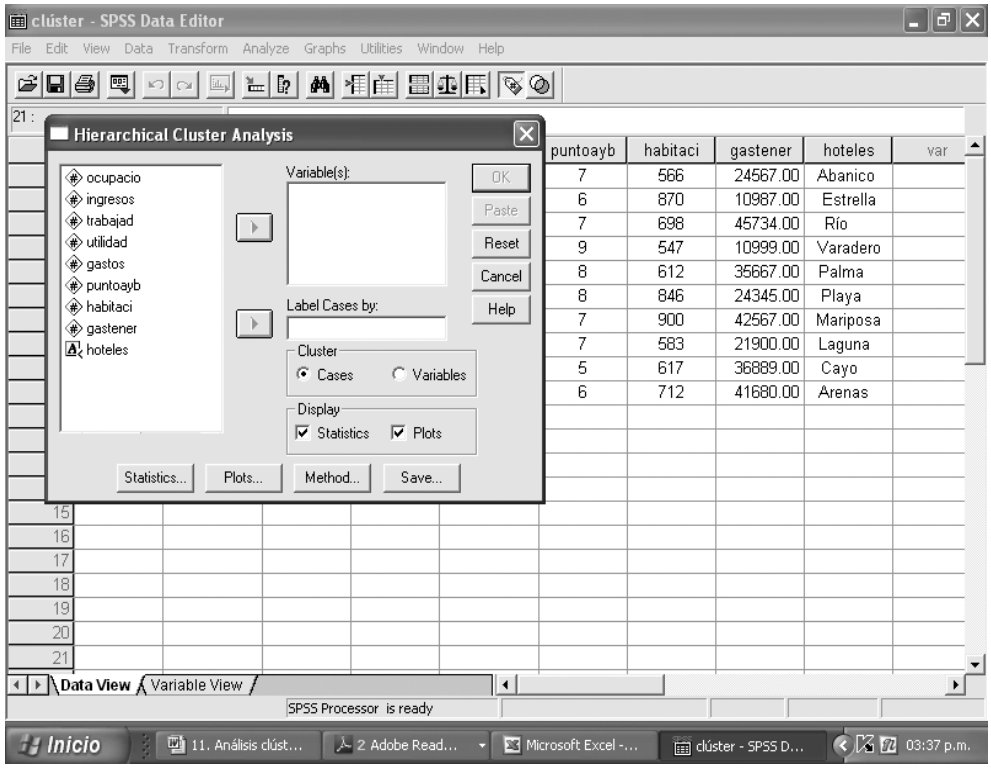
Hoteles	% ocupacio	ingresos	trabajad	utilidad	gastos	punto a+b	habita	gastener
Sirenis Abanico de Coral	46	101564.00	560	2463.00	99101.00	7	566	24567.00
Meliá Estrella de Mar	78	57890.00	315	1800.00	56090.00	6	870	10987.00
Iberostar Río Azul	65	114362.00	643	3101.00	111261.00	7	698	45734.00
Riu Varadero	59	87765.00	389	2746.00	85019.00	9	547	10999.00
Tryp Palma Real	74	103890.00	472	2834.00	101056.00	8	612	35667.00
Iberostar Playa Azul	49	92345.00	518	1964.00	90381.00	8	846	24345.00
Paradisus Mariposa Blanca	52	110321.00	589	946.00	109375.00	7	900	42567.00
Oasis Laguna Azul	66	74678.00	471	3123.00	71555.00	7	583	21900.00
Sol Cayo de Oro	91	98876.00	331	2680.00	96196.00	5	617	36889.00
Sandals Arenas	82	104564.00	470	1970.00	102594.00	6	712	41680.00

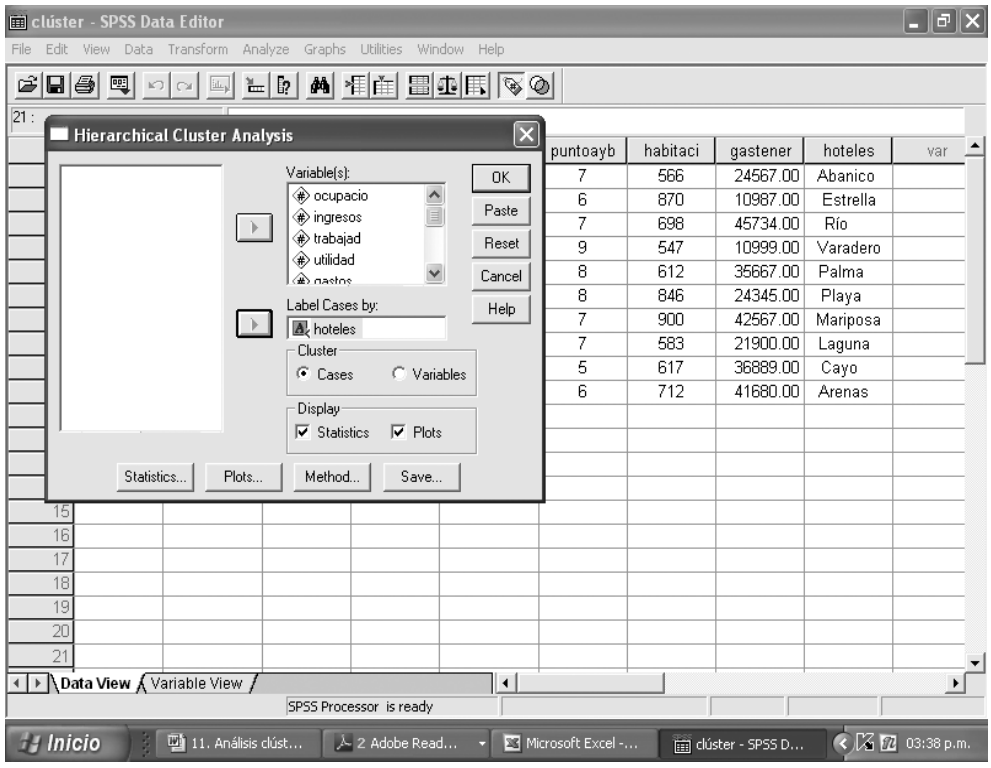
Solución:

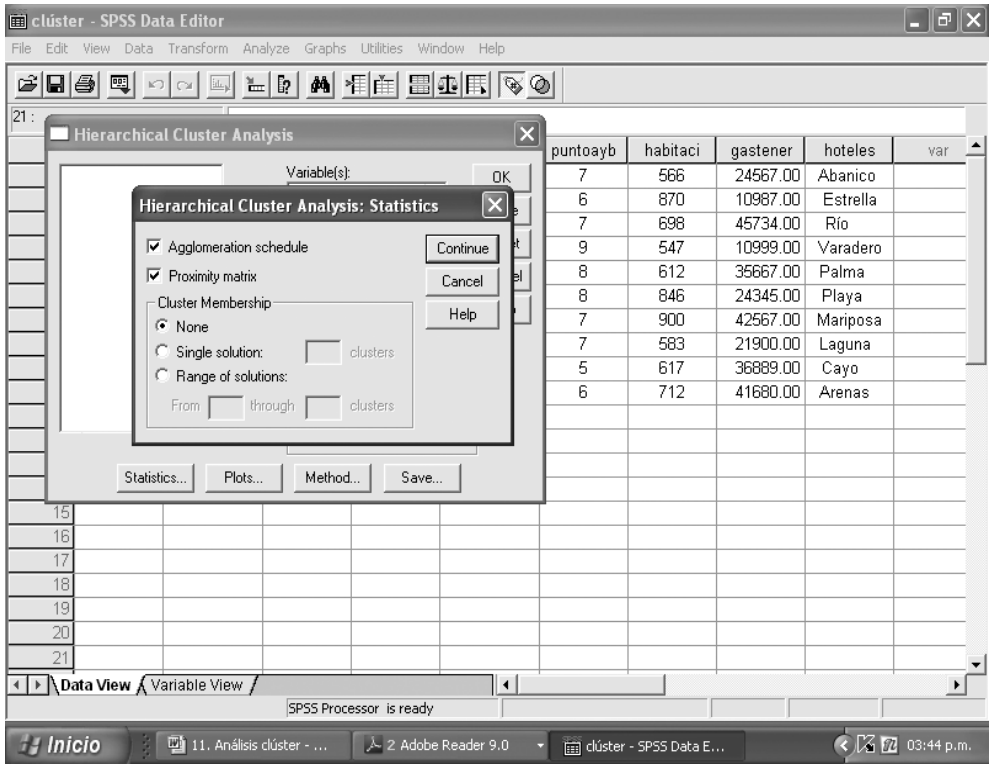
Empleando el SPSS, sería:

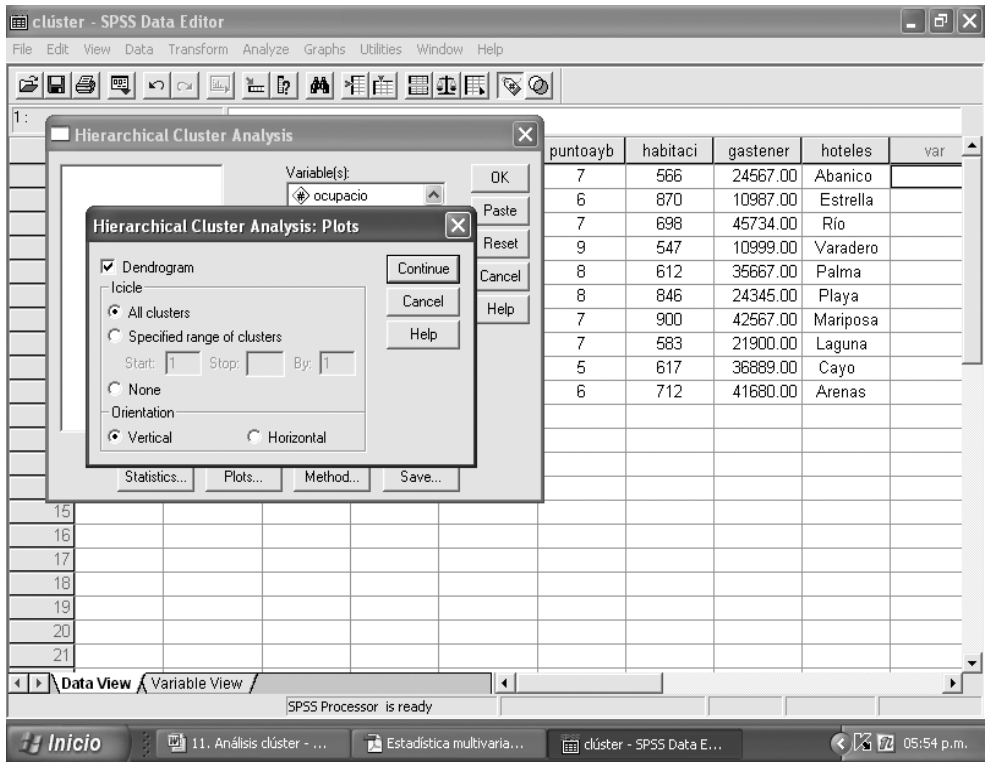

clúster - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
21 :
ocupacio ingresos trabajad utilidad gastos puntoayb habitaci gastener hoteles var
1 46 101564.00 560 2463.00 99101.00 7 566 24567.00 Abanico
2 78 57890.00 315 1800.00 56090.00 6 870 10987.00 Estrella
3 65 114362.00 643 3101.00 111261.00 7 698 45734.00 Río
4 59 87765.00 389 2746.00 85019.00 9 547 10999.00 Varadero
5 74 103890.00 472 2834.00 101056.00 8 612 35667.00 Palma
6 49 92345.00 518 1964.00 90381.00 8 846 24345.00 Playa
7 52 110321.00 589 946.00 109375.00 7 900 42567.00 Mariposa
8 66 74678.00 471 3123.00 71555.00 7 583 21900.00 Laguna
9 91 98876.00 331 2680.00 96196.00 5 617 36889.00 Cayo
10 82 104564.00 470 1970.00 102594.00 6 712 41680.00 Arenas
Data View Variable View
SPSS Processor is ready
Inicio 11. Análisis clúst... 2 Adobe Read... Microsoft Excel -... clúster - SPSS D... 03:34 p.m.

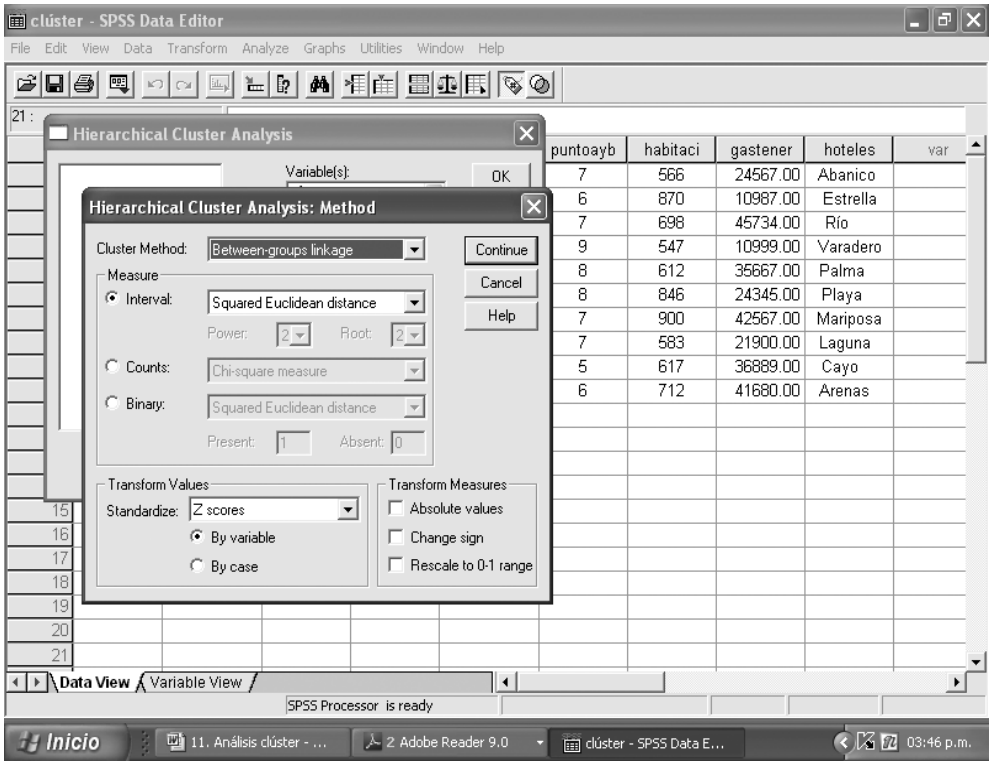

clúster - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
TwoStep Cluster...
K-Means Cluster...
Hierarchical Cluster...
Discriminant...
21 :
ocupacio ingreso gastos puntoayb habitaci gastener hoteles var
1 46 101564 99101.00 7 566 24567.00 Abanico
2 78 57890 56090.00 6 870 10987.00 Estrella
3 65 114362 111261.00 7 698 45734.00 Río
4 59 87765 85019.00 9 547 10999.00 Varadero
5 74 103890 612 35667.00 Palma
6 49 92345 846 24345.00 Playa
7 52 110321 900 42567.00 Mariposa
8 66 74678 583 21900.00 Laguna
9 91 98876 96196.00 5 617 36889.00 Cayo
10 82 104564 102594.00 6 712 41680.00 Arenas
Data View Variable View
Hierarchical Cluster
SPSS Processor is ready
Inicio 11. Análisis clúst... 2 Adobe Read... Microsoft Excel -... clúster - SPSS D... 03:37 p.m.

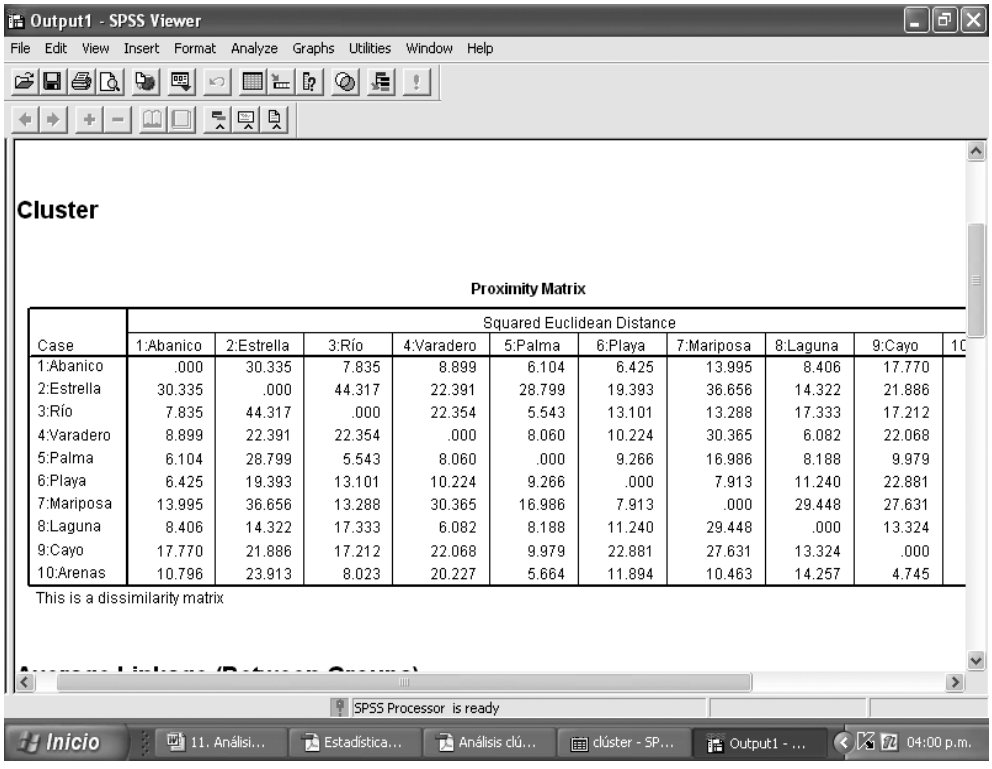

clúster - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Hierarchical Cluster Analysis
ocupacio
ingresos
trabajad
utilidad
gastos
puntoayb
habitaci
gastener
hoteles
Variable(s):
Label Cases by:
Cluster
Cases
Variables
Display
Statistics
Plots
OK
Paste
Reset
Cancel
Help
Statistics...
Plots...
Method...
Save...
puntoayb habitaci gastener hoteles var
7 566 24567.00 Abanico
6 870 10987.00 Estrella
7 698 45734.00 Río
9 547 10999.00 Varadero
8 612 35667.00 Palma
8 846 24345.00 Playa
7 900 42567.00 Mariposa
7 583 21900.00 Laguna
5 617 36889.00 Cayo
6 712 41680.00 Arenas
Data View
Variable View
SPSS Processor is ready
Inicio
11. Análisis clúst...
2 Adobe Read...
Microsoft Excel -...
clúster - SPSS D...
03:37 p.m.


clúster - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Hierarchical Cluster Analysis
Variable(s):
ocupacio
ingresos
trabajad
utilidad
Label Cases by:
hoteles
Cluster
Cases
Variables
Display
Statistics
Plots
OK
Paste
Reset
Cancel
Help
Statistics...
Plots...
Method...
Save...
puntoayb habitaci gastener hoteles var
7 566 24567.00 Abanico
6 870 10987.00 Estrella
7 698 45734.00 Río
9 547 10999.00 Varadero
8 612 35667.00 Palma
8 846 24345.00 Playa
7 900 42567.00 Mariposa
7 583 21900.00 Laguna
5 617 36889.00 Cayo
6 712 41680.00 Arenas
Data View
Variable View
SPSS Processor is ready
Inicio
11. Análisis clúst...
2 Adobe Read...
Microsoft Excel -...
clúster - SPSS D...
03:38 p.m.

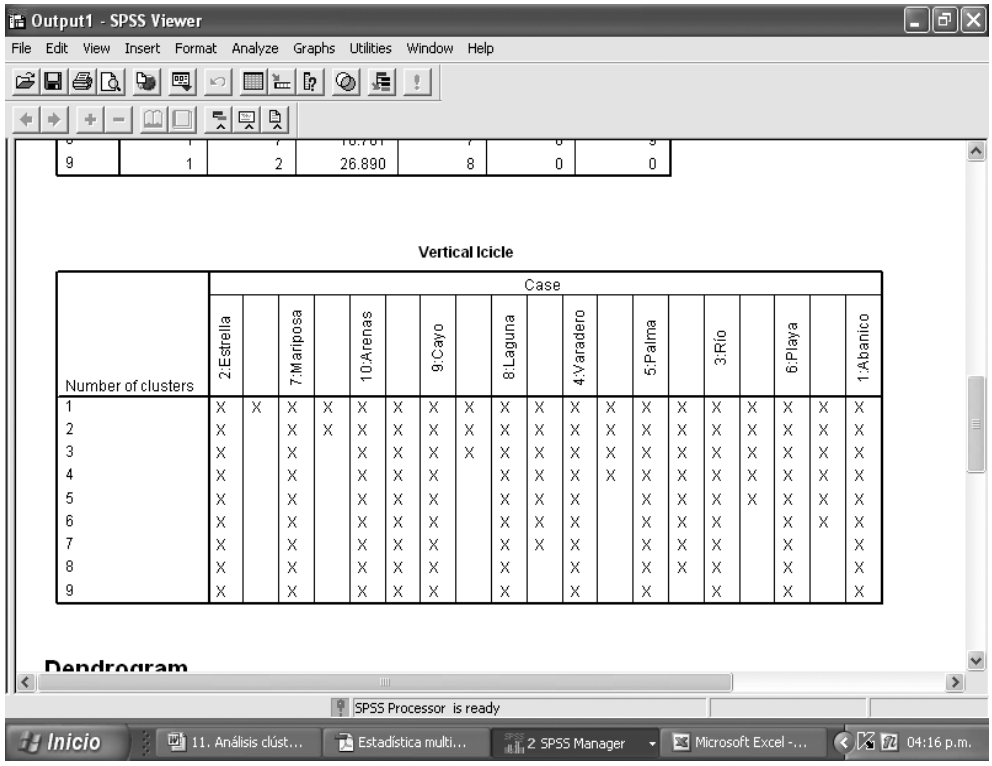

clúster - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Hierarchical Cluster Analysis
Variable(s):
OK
Hierarchical Cluster Analysis: Statistics
Agglomeration schedule
Proximity matrix
Cluster Membership
None
Single solution: clusters
Range of solutions:
From through clusters
Continue
Cancel
Help
Statistics... Plots... Method... Save...
puntoayb habitaci gastener hoteles var
7 566 24567.00 Abanico
6 870 10987.00 Estrella
7 698 45734.00 Río
9 547 10999.00 Varadero
8 612 35667.00 Palma
8 846 24345.00 Playa
7 900 42567.00 Mariposa
7 583 21900.00 Laguna
5 617 36889.00 Cayo
6 712 41680.00 Arenas
Data View Variable View
SPSS Processor is ready
Inicio
11. Análisis clúster - ...
2 Adobe Reader 9.0
clúster - SPSS Data E...
03:44 p.m.


clúster - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Hierarchical Cluster Analysis
Variable(s):
ocupacio
OK
Paste
Reset
Cancel
Help
Hierarchical Cluster Analysis: Plots
Dendrogram
Icicle
All clusters
Specified range of clusters
Start: Stop: By:
None
Orientation
Vertical Horizontal
Continue
Cancel
Help
Statistics... Plots... Method... Save...
puntoayb habitaci gastener hoteles var
7 566 24567.00 Abanico
6 870 10987.00 Estrella
7 698 45734.00 Río
9 547 10999.00 Varadero
8 612 35667.00 Palma
8 846 24345.00 Playa
7 900 42567.00 Mariposa
7 583 21900.00 Laguna
5 617 36889.00 Cayo
6 712 41680.00 Arenas
Data View Variable View
SPSS Processor is ready
Inicio
11. Análisis clúster - ...
Estadística multivaria...
clúster - SPSS Data E...
05:54 p.m.


clúster - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Hierarchical Cluster Analysis
Variable(s):
OK
Hierarchical Cluster Analysis: Method
Cluster Method: Between-groups linkage
Continue
Cancel
Help
Measure
Interval: Squared Euclidean distance
Power: 2 Root: 2
Counts: Chi-square measure
Binary: Squared Euclidean distance
Present: 1 Absent: 0
Transform Values
Standardize: Z scores
By variable
By case
Transform Measures
Absolute values
Change sign
Rescale to 0-1 range
puntoayb habitaci gastener hoteles var
7 566 24567.00 Abanico
6 870 10987.00 Estrella
7 698 45734.00 Río
9 547 10999.00 Varadero
8 612 35667.00 Palma
8 846 24345.00 Playa
7 900 42567.00 Mariposa
7 583 21900.00 Laguna
5 617 36889.00 Cayo
6 712 41680.00 Arenas
Data View Variable View
SPSS Processor is ready
Inicio
11. Análisis clúster - ...
2 Adobe Reader 9.0
clúster - SPSS Data E...
03:46 p.m.


Output1 - SPSS Viewer
File Edit View Insert Format Analyze Graphs Utilities Window Help
Cluster
Proximity Matrix
Squared Euclidean Distance
Case 1:Abanico 2:Estrella 3:Río 4:Varadero 5:Palma 6:Playa 7:Mariposa 8:Laguna 9:Cayo
1:Abanico .000 30.335 7.835 8.899 6.104 6.425 13.995 8.406 17.770
2:Estrella 30.335 .000 44.317 22.391 28.799 19.393 36.656 14.322 21.886
3:Río 7.835 44.317 .000 22.354 5.543 13.101 13.288 17.333 17.212
4:Varadero 8.899 22.391 22.354 .000 8.060 10.224 30.365 6.082 22.068
5:Palma 6.104 28.799 5.543 8.060 .000 9.266 16.986 8.188 9.979
6:Playa 6.425 19.393 13.101 10.224 9.266 .000 7.913 11.240 22.881
7:Mariposa 13.995 36.656 13.288 30.365 16.986 7.913 .000 29.448 27.631
8:Laguna 8.406 14.322 17.333 6.082 8.188 11.240 29.448 .000 13.324
9:Cayo 17.770 21.886 17.212 22.068 9.979 22.881 27.631 13.324 .000
10:Arenas 10.796 23.913 8.023 20.227 5.664 11.894 10.463 14.257 4.745
This is a dissimilarity matrix
SPSS Processor is ready
Inicio
11. Análisi...
Estadística...
Análisis clú...
clúster - SP...
Output1 - ...
04:00 p.m.

En la imagen anterior, se observa la tabla "*Proximity Matrix*" donde se muestran los coeficientes de distancia euclídea al cuadrado, entre los distintos hoteles de la muestra. Por ejemplo, la distancia o diferencia mayor con un coeficiente igual a 44.317, es la existente entre los hoteles Iberostar Río Azul y Meliá Estrella de Mar. Por el contrario, los más próximos o parecidos, son los hoteles Sol Cayo de Oro y Sandals Arenas con un coeficiente igual a 4.745.



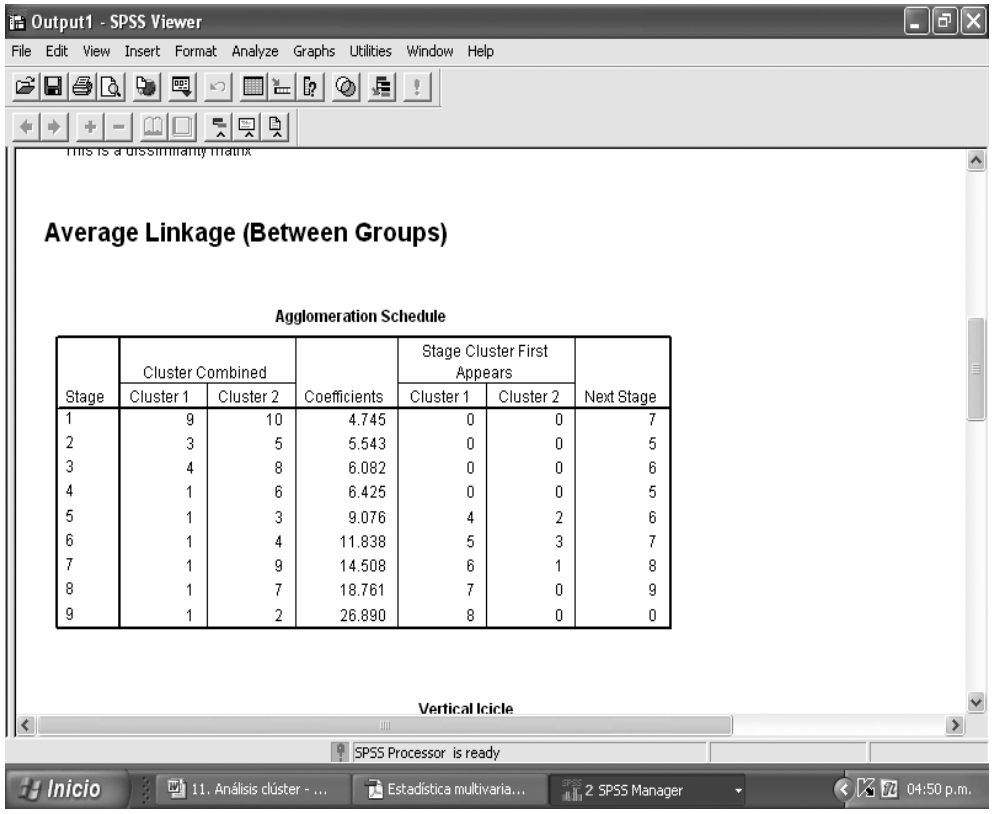
Number of clusters	2:Estrella		7:Mariposa		10:Arenas		9:Cayo		8:Laguna		4:Varadero		5:Palma		3:Río		6:Playa		1:Abanico
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
3	X		X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
4	X		X		X	X	X		X	X	X	X	X	X	X	X	X	X	X
5	X		X		X	X	X		X	X	X		X	X	X	X	X	X	X
6	X		X		X	X	X		X	X	X		X	X	X		X	X	X
7	X		X		X	X	X		X	X	X		X	X	X		X		X
8	X		X		X	X	X		X		X		X	X	X		X		X
9	X		X		X	X	X		X		X		X		X		X		X

En la imagen anterior, se muestra el gráfico de carámbanos o tabla "*Vertical Icicle*" donde se puede ir determinando los diferentes clusters a cada nivel. Obsérvese que el:

- primer cluster está formado por los hoteles Sandals Arenas y Sol Cayo de Oro
- segundo cluster: Tryp Palma Real e Iberostar Río Azul
- tercer cluster: Oasis Laguna Azul y Riu Varadero
- cuarto cluster: Iberostar Playa Azul y Sirenis Abanico de Coral
- quinto cluster (primer multicluster): Tryp Palma Real, Iberostar Río Azul,

Iberostar Playa Azul y Sirenis Abanico de Coral

- sexto cluster (segundo multicluster): Oasis Laguna Azul, Riu Varadero, Tryp Palma Real, Iberostar Río Azul, Iberostar Playa Azul y Sirenis Abanico de Coral
- séptimo cluster (tercer multicluster): Sandals Arenas, Sol Cayo de Oro, Oasis Laguna Azul, Riu Varadero, Tryp Palma Real, Iberostar Río Azul, Iberostar Playa Azul y Sirenis Abanico de Coral
- octavo cluster (cuarto multicluster): Paradisus Mariposa Blanca, Sandals Arenas, Sol Cayo de Oro, Oasis Laguna Azul, Riu Varadero, Tryp Palma Real, Iberostar Río Azul, Iberostar Playa Azul y Sirenis Abanico de Coral
- noveno cluster: incluye todos los hoteles



Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	9	10	4.745	0	0	7
2	3	5	5.543	0	0	5
3	4	8	6.082	0	0	6
4	1	6	6.425	0	0	5
5	1	3	9.076	4	2	6
6	1	4	11.838	5	3	7
7	1	9	14.508	6	1	8
8	1	7	18.761	7	0	9
9	1	2	26.890	8	0	0

En la imagen anterior, se muestra la tabla "*Agglomeration Schedule*" donde se observa que, por ejemplo, en el primer nivel se unen para formar un cluster, los hoteles 9 (Sol Cayo de Oro) y 10 (Sandals Arenas). Ambos casos se unen a otros hoteles por primera vez para formar un multicluster, en el nivel siete,

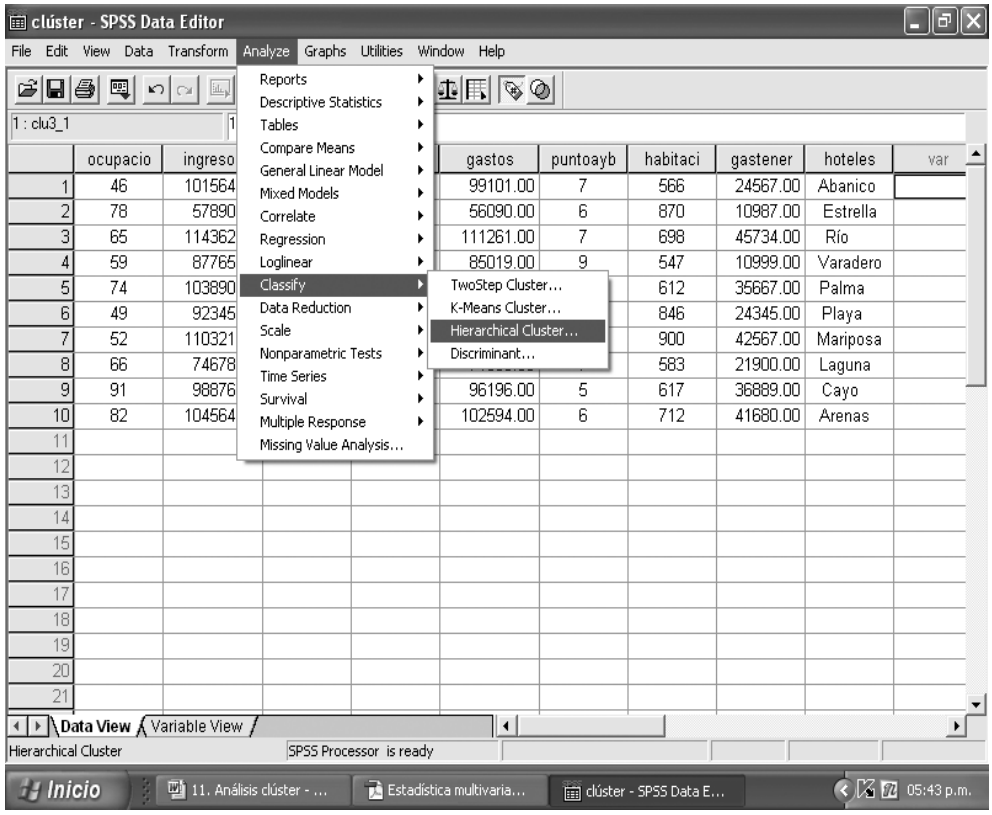
cuando se les suma los hoteles Oasis Laguna Azul, Riu Varadero, Tryp Palma Real, Iberostar Río Azul, Iberostar Playa Azul y Sirenis Abanico de Coral.

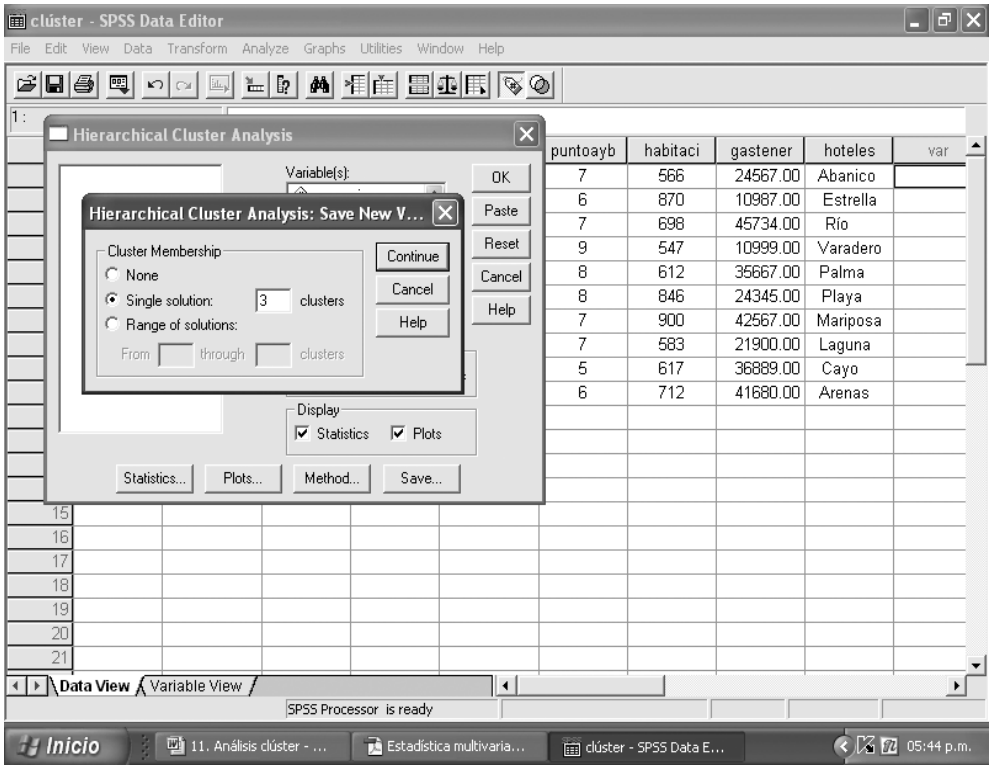
En esta misma tabla se observa el valor del coeficiente para cada nivel, de modo que mientras menor sea el coeficiente, indicará la existencia de clusters más homogéneos. Cuanto mayor sea el valor del coeficiente, pues más heterogéneos serán éstos.

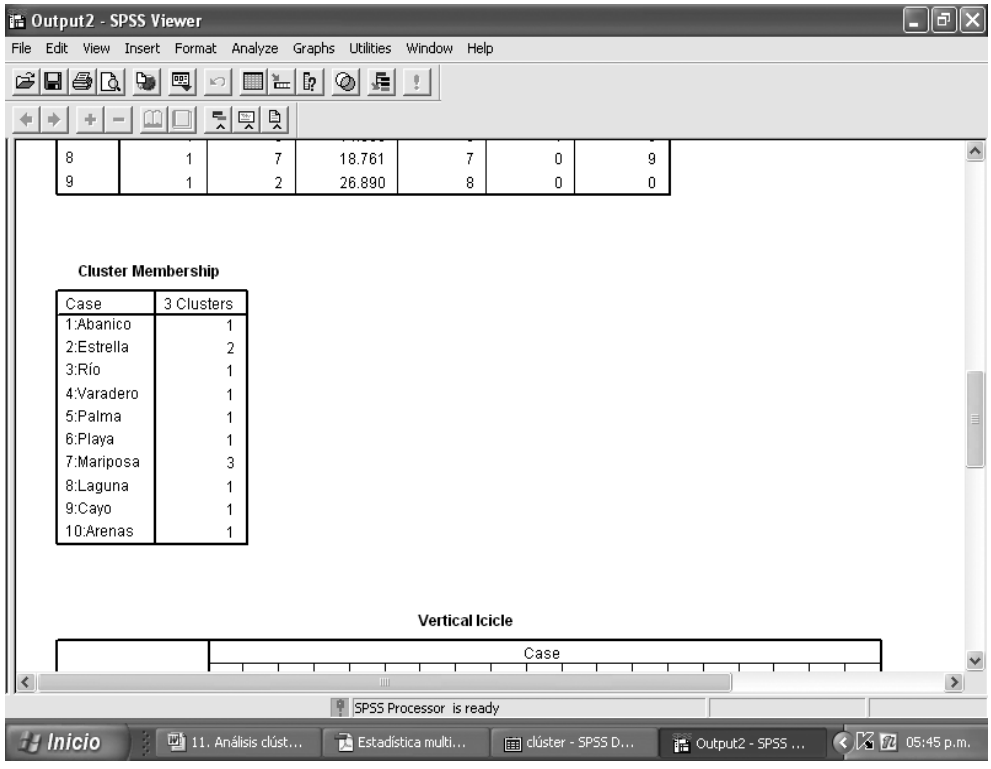
Supóngase que el grupo de analistas de la Delegación del MINTUR, desea obtener una cantidad de clusters específicos de la muestra de hoteles tomada, en este caso, 3 clusters.

Solución:

Empleando el SPSS, sería:

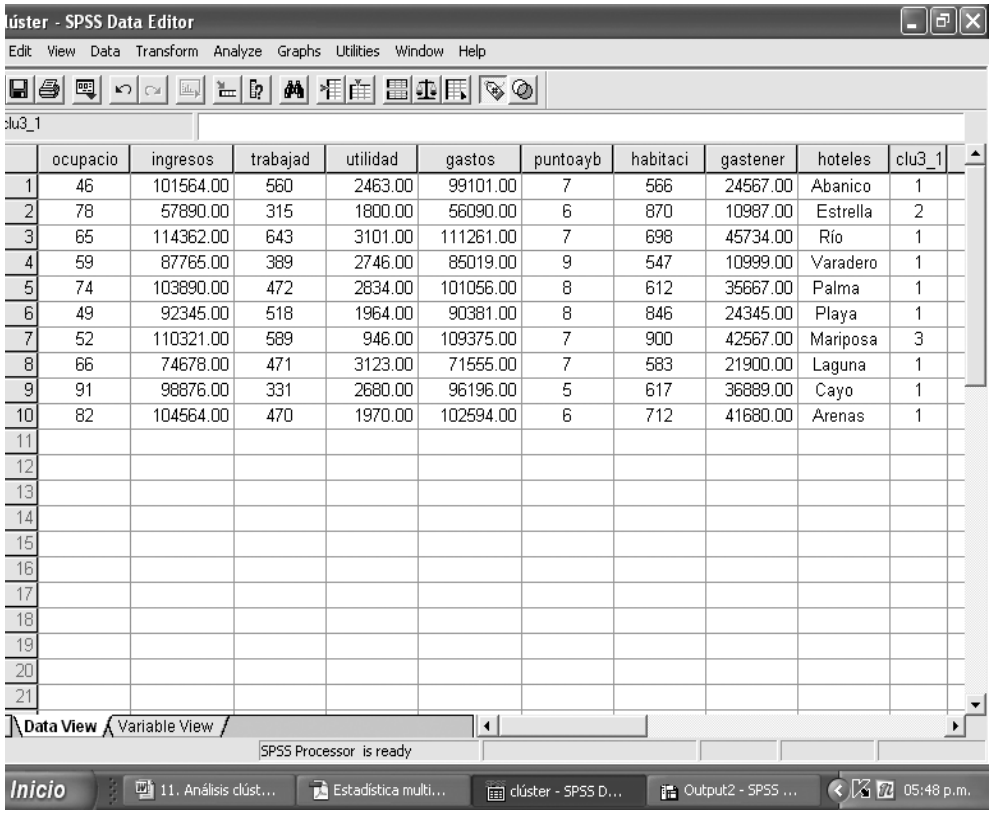



clúster - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Hierarchical Cluster Analysis
Variable(s):
OK
Paste
Reset
Cancel
Help
Hierarchical Cluster Analysis: Save New V...
Cluster Membership
None
Single solution: 3 clusters
Range of solutions:
From through clusters
Continue
Cancel
Help
Display
Statistics
Plots
Statistics...
Plots...
Method...
Save...
puntoayb habitaci gastener hoteles var
7 566 24567.00 Abanico
6 870 10987.00 Estrella
7 698 45734.00 Río
9 547 10999.00 Varadero
8 612 35667.00 Palma
8 846 24345.00 Playa
7 900 42567.00 Mariposa
7 583 21900.00 Laguna
5 617 36889.00 Cayo
6 712 41680.00 Arenas
Data View
Variable View
SPSS Processor is ready
Inicio
11. Análisis clúster - ...
Estadística multivaria...
clúster - SPSS Data E...
05:44 p.m.


Output2 - SPSS Viewer
File Edit View Insert Format Analyze Graphs Utilities Window Help
8 1 7 18.761 7 0 9
9 1 2 26.890 8 0 0
Cluster Membership
Case 3 Clusters
1:Abanico 1
2:Estrella 2
3:Río 1
4:Varadero 1
5:Palma 1
6:Playa 1
7:Mariposa 3
8:Laguna 1
9:Cayo 1
10:Arenas 1
Vertical Icicle
Case
SPSS Processor is ready
Inicio
11. Análisis clúst...
Estadística multi...
clúster - SPSS D...
Output2 - SPSS ...
05:45 p.m.

En la imagen anterior, se muestra la tabla "*Cluster Membership*" donde se observa que el:

- primer cluster está formado por los hoteles: Sirenis Abanico de Coral, Iberostar Río Azul, Riu Varadero, Tryp Palma Real, Iberostar Playa Azul, Oasis Laguna Azul, Sol Cayo de Oro y Sandals Arenas
- segundo cluster: Meliá Estrella de Mar
- tercer cluster: Paradisus Mariposa Blanca



	ocupacio	ingresos	trabajad	utilidad	gastos	puntoayb	habitaci	gastener	hoteles	clu3_1
1	46	101564.00	560	2463.00	99101.00	7	566	24567.00	Abanico	1
2	78	57890.00	315	1800.00	56090.00	6	870	10987.00	Estrella	2
3	65	114362.00	643	3101.00	111261.00	7	698	45734.00	Río	1
4	59	87765.00	389	2746.00	85019.00	9	547	10999.00	Varadero	1
5	74	103890.00	472	2834.00	101056.00	8	612	35667.00	Palma	1
6	49	92345.00	518	1964.00	90381.00	8	846	24345.00	Playa	1
7	52	110321.00	589	946.00	109375.00	7	900	42567.00	Mariposa	3
8	66	74678.00	471	3123.00	71555.00	7	583	21900.00	Laguna	1
9	91	98876.00	331	2680.00	96196.00	5	617	36889.00	Cayo	1
10	82	104564.00	470	1970.00	102594.00	6	712	41680.00	Arenas	1
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										

Obsérvese en la imagen anterior, que a la base de datos original, el programa ha añadido una nueva columna llamada "clu3_1". La misma refleja igual contenido que el de la tabla "*Cluster Membership*" analizada previamente, o sea, a qué cluster pertenece cada hotel dado que han sido seleccionados 3 clusters.

Ahora véase un ejemplo de análisis cluster K-medias.

Ejemplo 2:

El grupo de analistas de la Delegación del MINTUR, ha decidido ahora ampliar la muestra de hoteles a estudiar a treinta y tres. Continuando el análisis de las ocho variables en las entidades hoteleras, estas últimas se mencionan a continuación:

Hoteles	% ocupacio	ingresos	trabajad	utilidad	gastos	punto a+b	habita.	gastener
Sirenis Abanico de Coral	46	10156.00	560	2463.00	99101.00	7	566	24567.00
Meliá Estrella de Mar	78	57890.00	315	1800.00	56090.00	6	870	10987.00
Iberostar Río Azul	65	114362.00	643	3101.00	111261.00	7	698	45734.00
Riu Varadero	59	87765.00	389	2746.00	85019.00	9	547	10999.00
Tryp Palma Real	74	103890.00	472	2834.00	101056.00	8	612	35667.00
Iberostar Playa Azul	49	92345.00	518	1964.00	90381.00	8	846	24345.00
Paradisus Mariposa Blanca	52	110321.00	589	946.00	109375.00	7	900	42567.00
Oasis Laguna Azul	66	74678.00	471	3123.00	71555.00	7	583	21900.00
Sol Cayo de Oro	91	98876.00	331	2680.00	96196.00	5	617	36889.00
Sandals Arenas	82	104564.00	470	1970.00	102594.00	6	712	41680.00
Iberostar Princesa Roja	49	68014.00	396	1025.00	79653.00	6	612	10258.00

Meliá Fuertes Vientos	58	102563.00	525	987.00	102589.00	7	745	26985.00
Riu Piedra Dorada	46	95562.00	654	2589.00	98563.00	8	896	31489.00
Iberostar Orilla Azul	61	79586.00	489	3125.00	84125.00	6	625	41288.00
Sirenis Coral de Fuego	72	100589.00	369	985.00	98745.00	7	596	32589.00
Oasis Caleta Buena	83	46892.00	489	3152.00	100258.00	8	910	21478.00
Meliá Aguas Claras	69	95411.00	526	1489.00	78965.00	8	856	10256.00
Sol Lago Azul	58	78589.00	621	2541.00	82336.00	7	678	25963.00
Tryp Luna Plateada	61	74258.00	368	3214.00	97268.00	9	748	32589.00
Meliá Duna Alta	84	96236.00	489	1025.00	100569.00	7	908	45632.00
Paradisus Tocororo	91	85555.00	526	2365.00	98425.00	6	658	25825.00
Iberostar Las Morlas	54	84259.00	514	1478.00	84856.00	8	547	14785.00
Tryp Cielo Azul	51	79463.00	621	2589.00	71937.00	6	963	30156.00
Villa Real	49	68954.00	358	3654.00	100485.00	8	852	21485.00
Riu Mar Profundo	66	81258.00	369	1485.00	10632.00	8	741	10325.00
Paradisus Patriarca	74	79589.00	321	965.00	89652.00	6	789	26985.00
Lago Verde	89	102596.00	412	987.00	74589.00	9	654	45698.00
Sol Palacio	91	105478.00	562	2589.00	96541.00	7	523	32156.00
Barceló Sol Brillante	65	84563.00	458	954.00	100256.00	6	789	12589.00
Oasis Canal Grande	48	100892.00	363	1236.00	96369.00	8	954	23589.00
Pino Alto	58	98456.00	458	2563.00	89652.00	8	741	10258.00

Sirenis Esponja de Mar	64	101548.00	589	3111.00	79658.00	7	852	12365.00
Playa Larga	71	78963.00	612	1025.00	89654.00	6	693	14859.00

Basándose en los datos recopilados de ocho variables que han sido medidas en cada uno de los treinta y tres hoteles, el objetivo de los miembros del grupo, consiste en agrupar dichas entidades según su similitud o semejanza, pero predeterminando la cantidad de clusters a 4.

Solución:

Empleando el SPSS, sería:

clúster - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

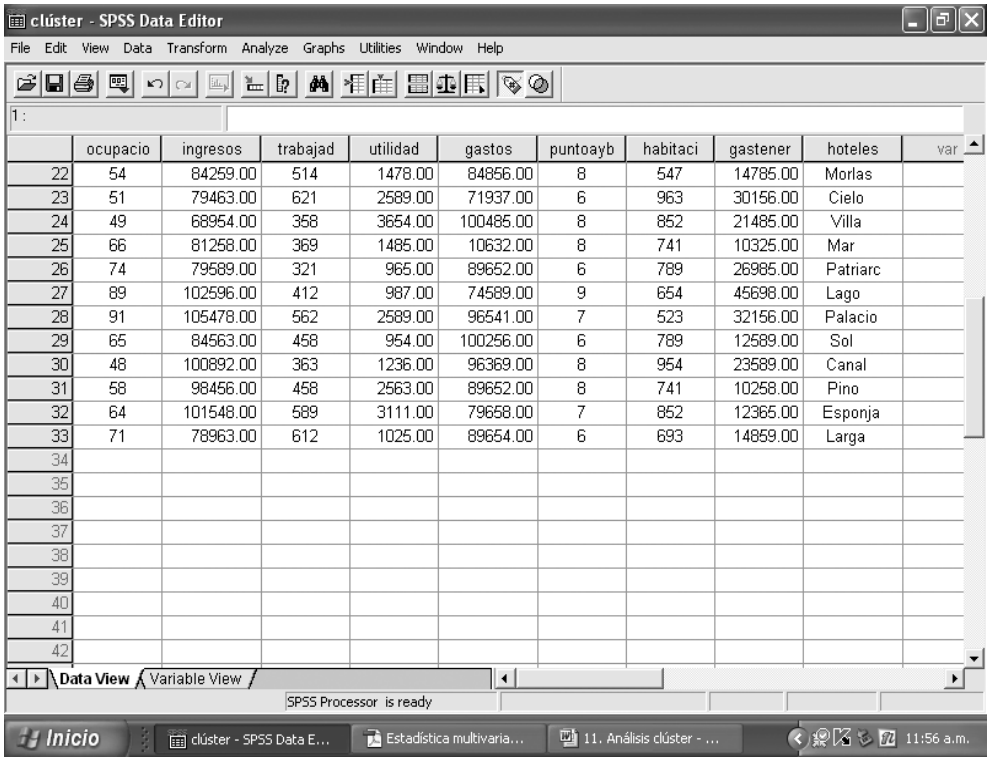
1 :

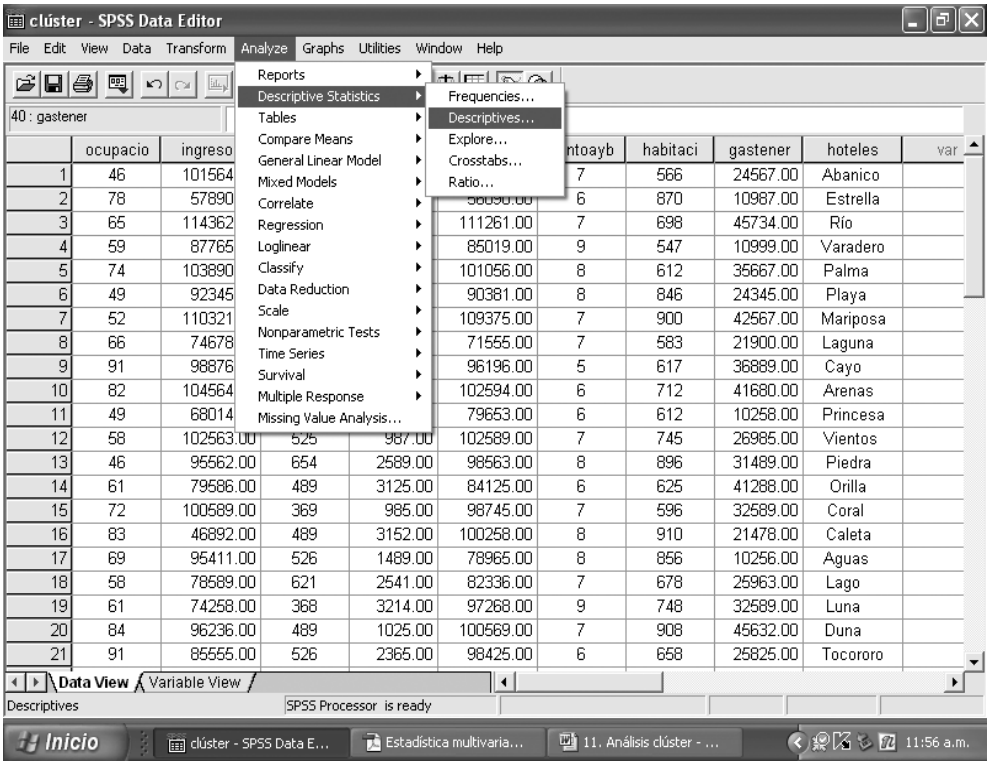
	ocupacio	ingresos	trabajad	utilidad	gastos	puntoayb	habitaci	gastener	hoteles
1	46	101564.00	560	2463.00	99101.00	7	566	24567.00	Abanico
2	78	57890.00	315	1800.00	56090.00	6	870	10987.00	Estrella
3	65	114362.00	643	3101.00	111261.00	7	698	45734.00	Río
4	59	87765.00	389	2746.00	85019.00	9	547	10999.00	Varadero
5	74	103890.00	472	2834.00	101056.00	8	612	35667.00	Palma
6	49	92345.00	518	1964.00	90381.00	8	846	24345.00	Playa
7	52	110321.00	589	946.00	109375.00	7	900	42567.00	Mariposa
8	66	74678.00	471	3123.00	71555.00	7	583	21900.00	Laguna
9	91	98876.00	331	2680.00	96196.00	5	617	36889.00	Cayo
10	82	104564.00	470	1970.00	102594.00	6	712	41680.00	Arenas
11	49	68014.00	396	1025.00	79653.00	6	612	10258.00	Princesa
12	58	102563.00	525	987.00	102589.00	7	745	26985.00	Vientos
13	46	95562.00	654	2589.00	98563.00	8	896	31489.00	Piedra
14	61	79586.00	489	3125.00	84125.00	6	625	41288.00	Orilla
15	72	100589.00	369	985.00	98745.00	7	596	32589.00	Coral
16	83	46892.00	489	3152.00	100258.00	8	910	21478.00	Caleta
17	69	95411.00	526	1489.00	78965.00	8	856	10256.00	Aguas
18	58	78589.00	621	2541.00	82336.00	7	678	25963.00	Lago
19	61	74258.00	368	3214.00	97268.00	9	748	32589.00	Luna
20	84	96236.00	489	1025.00	100569.00	7	908	45632.00	Duna
21	91	85555.00	526	2365.00	98425.00	6	658	25825.00	Tocororo

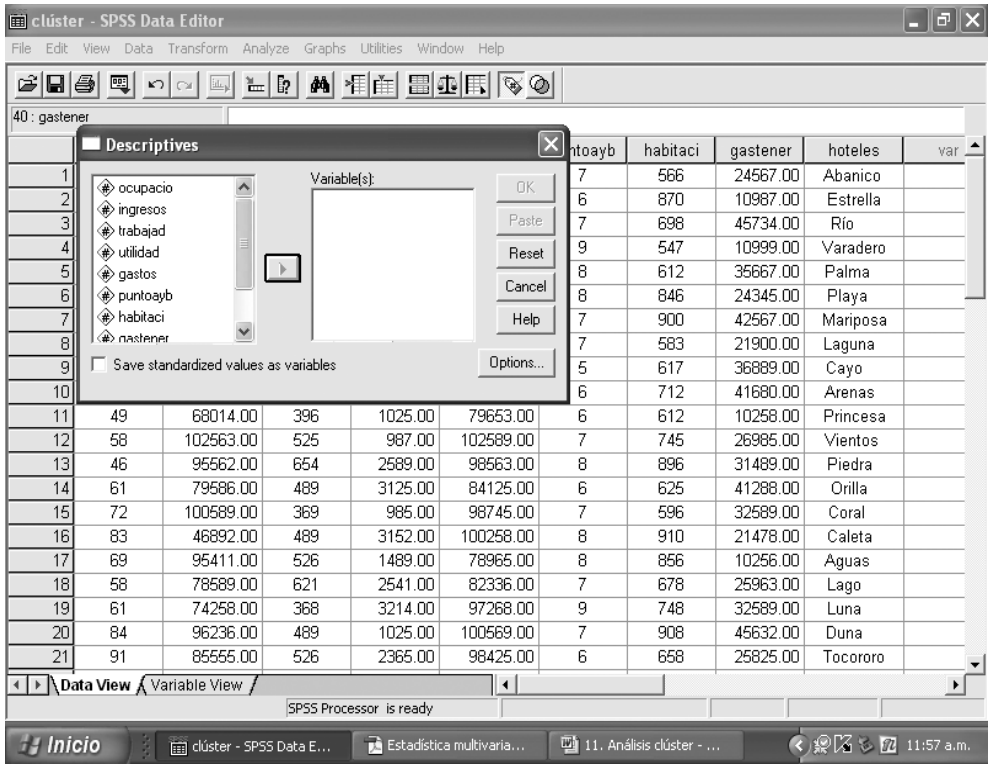
Data View | Variable View

SPSS Processor is ready

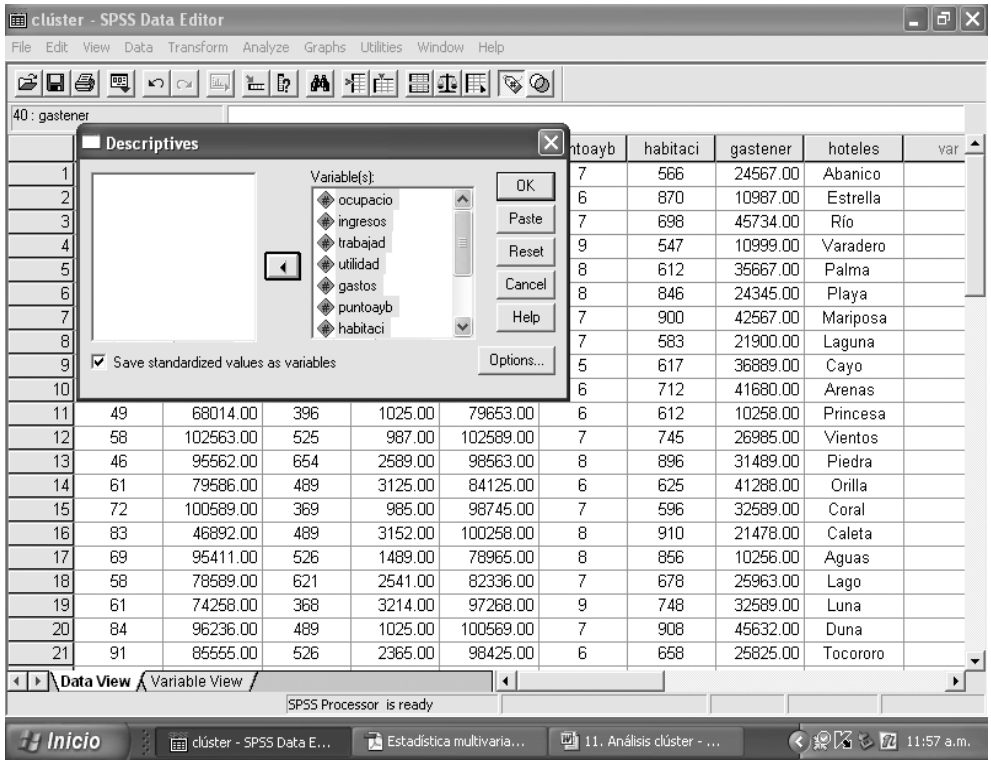
Inicio | clúster - SPSS Data E... | Estadística multivaria... | 11. Análisis clúster - ...


clúster - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 :
ocupacio ingresos trabajad utilidad gastos puntoayb habitaci gastener hoteles var
22 54 84259.00 514 1478.00 84856.00 8 547 14785.00 Morlas
23 51 79463.00 621 2589.00 71937.00 6 963 30156.00 Cielo
24 49 68954.00 358 3654.00 100485.00 8 852 21485.00 Villa
25 66 81258.00 369 1485.00 10632.00 8 741 10325.00 Mar
26 74 79589.00 321 965.00 89652.00 6 789 26985.00 Patriarc
27 89 102596.00 412 987.00 74589.00 9 654 45698.00 Lago
28 91 105478.00 562 2589.00 96541.00 7 523 32156.00 Palacio
29 65 84563.00 458 954.00 100256.00 6 789 12589.00 Sol
30 48 100892.00 363 1236.00 96369.00 8 954 23589.00 Canal
31 58 98456.00 458 2563.00 89652.00 8 741 10258.00 Pino
32 64 101548.00 589 3111.00 79658.00 7 852 12365.00 Esponja
33 71 78963.00 612 1025.00 89654.00 6 693 14859.00 Larga
Data View Variable View
SPSS Processor is ready
Inicio
clúster - SPSS Data E...
Estadística multivaria...
11. Análisis clúster - ...
11:56 a.m.

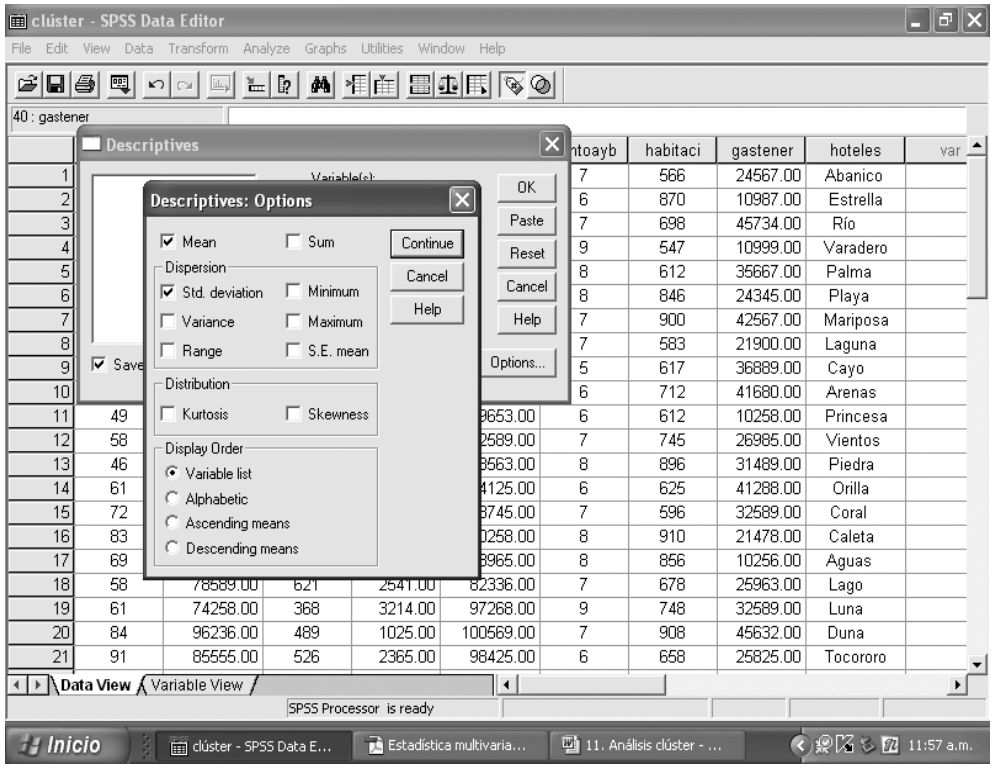

clúster - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Frequencies...
Descriptives...
Explore...
Crosstabs...
Ratio...
40 : gastener
ocupacio ingreso ntoayb habitaci gastener hoteles var
1 46 101564 7 566 24567.00 Abanico
2 78 57890 6 870 10987.00 Estrella
3 65 114362 111261.00 7 698 45734.00 Río
4 59 87765 85019.00 9 547 10999.00 Varadero
5 74 103890 101056.00 8 612 35667.00 Palma
6 49 92345 90381.00 8 846 24345.00 Playa
7 52 110321 109375.00 7 900 42567.00 Mariposa
8 66 74678 71555.00 7 583 21900.00 Laguna
9 91 98876 96196.00 5 617 36889.00 Cayo
10 82 104564 102594.00 6 712 41680.00 Arenas
11 49 68014 79653.00 6 612 10258.00 Princesa
12 58 102563.00 525 987.00 102589.00 7 745 26985.00 Vientos
13 46 95562.00 654 2589.00 98563.00 8 896 31489.00 Piedra
14 61 79586.00 489 3125.00 84125.00 6 625 41288.00 Orilla
15 72 100589.00 369 985.00 98745.00 7 596 32589.00 Coral
16 83 46892.00 489 3152.00 100258.00 8 910 21478.00 Caleta
17 69 95411.00 526 1489.00 78965.00 8 856 10256.00 Aguas
18 58 78589.00 621 2541.00 82336.00 7 678 25963.00 Lago
19 61 74258.00 368 3214.00 97268.00 9 748 32589.00 Luna
20 84 96236.00 489 1025.00 100569.00 7 908 45632.00 Duna
21 91 85555.00 526 2365.00 98425.00 6 658 25825.00 Tocororo
Data View Variable View
Descriptives
SPSS Processor is ready
Inicio
clúster - SPSS Data E...
Estadística multivaria...
11. Análisis clúster - ...
11:56 a.m.

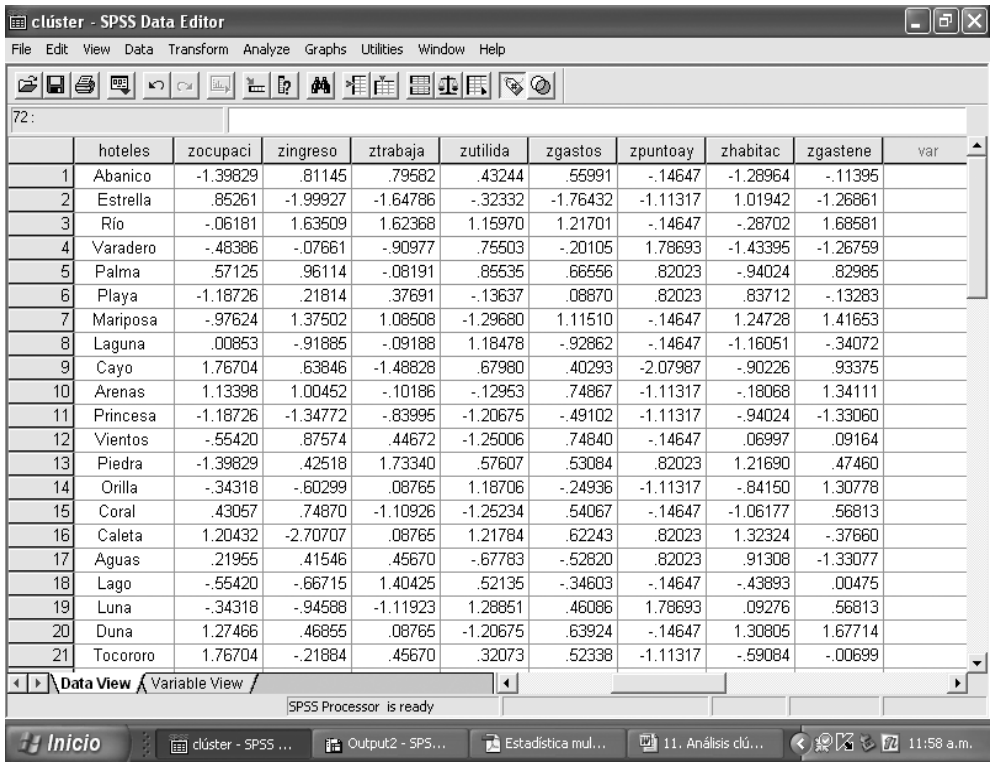


						ntoayb	habitaci	gastener	hoteles	var
1						7	566	24567.00	Abanico	
2						6	870	10987.00	Estrella	
3						7	698	45734.00	Río	
4						9	547	10999.00	Varadero	
5						8	612	35667.00	Palma	
6						8	846	24345.00	Playa	
7						7	900	42567.00	Mariposa	
8						7	583	21900.00	Laguna	
9						5	617	36889.00	Cayo	
10						6	712	41680.00	Arenas	
11	49	68014.00	396	1025.00	79653.00	6	612	10258.00	Princesa	
12	58	102563.00	525	987.00	102589.00	7	745	26985.00	Vientos	
13	46	95562.00	654	2589.00	98563.00	8	896	31489.00	Piedra	
14	61	79586.00	489	3125.00	84125.00	6	625	41288.00	Orilla	
15	72	100589.00	369	985.00	98745.00	7	596	32589.00	Coral	
16	83	46892.00	489	3152.00	100258.00	8	910	21478.00	Caleta	
17	69	95411.00	526	1489.00	78965.00	8	856	10256.00	Aguas	
18	58	78589.00	621	2541.00	82336.00	7	678	25963.00	Lago	
19	61	74258.00	368	3214.00	97268.00	9	748	32589.00	Luna	
20	84	96236.00	489	1025.00	100569.00	7	908	45632.00	Duna	
21	91	85555.00	526	2365.00	98425.00	6	658	25825.00	Tocororo	

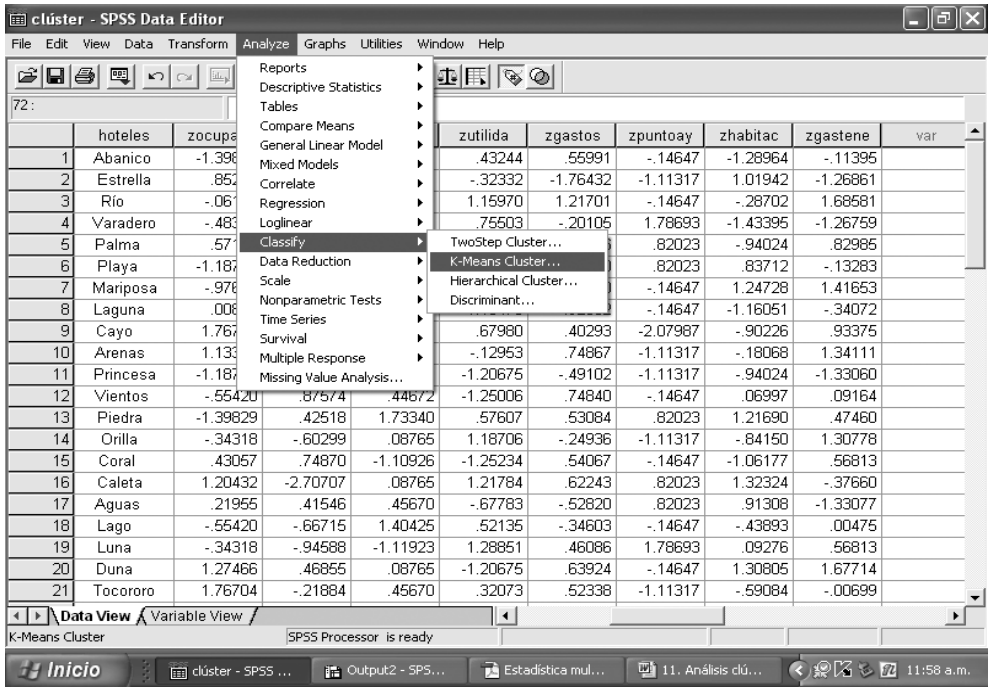


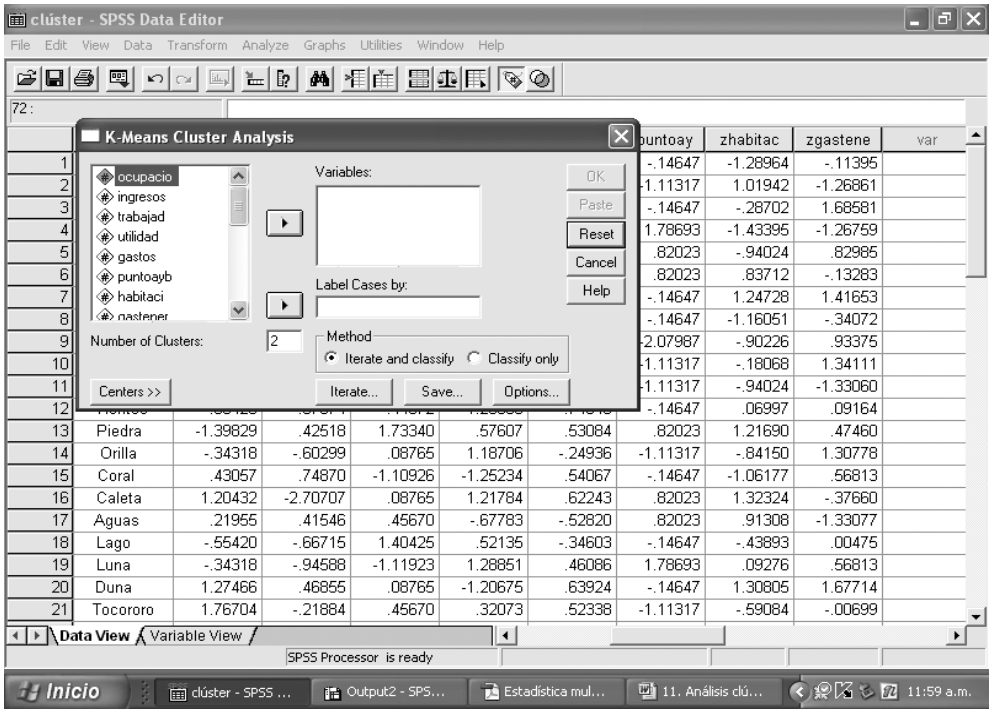
						ntoayb	habitaci	gastener	hoteles	var
1						7	566	24567.00	Abanico	
2						6	870	10987.00	Estrella	
3						7	698	45734.00	Río	
4						9	547	10999.00	Varadero	
5						8	612	35667.00	Palma	
6						8	846	24345.00	Playa	
7						7	900	42567.00	Mariposa	
8						7	583	21900.00	Laguna	
9						5	617	36889.00	Cayo	
10						6	712	41680.00	Arenas	
11	49	68014.00	396	1025.00	79653.00	6	612	10258.00	Princesa	
12	58	102563.00	525	987.00	102589.00	7	745	26985.00	Vientos	
13	46	95562.00	654	2589.00	98563.00	8	896	31489.00	Piedra	
14	61	79586.00	489	3125.00	84125.00	6	625	41288.00	Orilla	
15	72	100589.00	369	985.00	98745.00	7	596	32589.00	Coral	
16	83	46892.00	489	3152.00	100258.00	8	910	21478.00	Caleta	
17	69	95411.00	526	1489.00	78965.00	8	856	10256.00	Aguas	
18	58	78589.00	621	2541.00	82336.00	7	678	25963.00	Lago	
19	61	74258.00	368	3214.00	97268.00	9	748	32589.00	Luna	
20	84	96236.00	489	1025.00	100569.00	7	908	45632.00	Duna	
21	91	85555.00	526	2365.00	98425.00	6	658	25825.00	Tocororo	

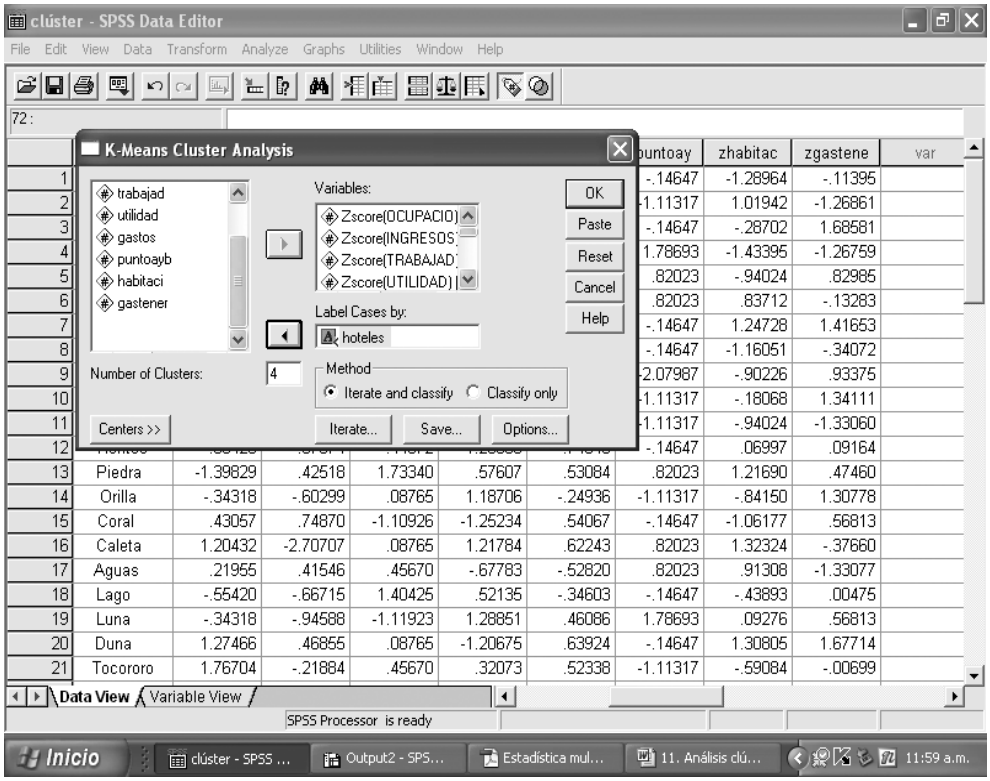

clúster - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
40 : gastener
Descriptives
Descriptives: Options
Mean
Sum
Dispersion
Std. deviation
Minimum
Variance
Maximum
Range
S.E. mean
Distribution
Kurtosis
Skewness
Display Order
Variable list
Alphabetic
Ascending means
Descending means
Continue
Cancel
Help
OK
Paste
Reset
Cancel
Help
Options...
habitaci gastener hoteles
566 24567.00 Abanico
870 10987.00 Estrella
698 45734.00 Río
547 10999.00 Varadero
612 35667.00 Palma
846 24345.00 Playa
900 42567.00 Mariposa
583 21900.00 Laguna
617 36889.00 Cayo
712 41680.00 Arenas
612 10258.00 Princesa
745 26985.00 Vientos
896 31489.00 Piedra
625 41288.00 Orilla
596 32589.00 Coral
910 21478.00 Caleta
856 10256.00 Aguas
678 25963.00 Lago
748 32589.00 Luna
908 45632.00 Duna
658 25825.00 Tocororo
Data View Variable View
SPSS Processor is ready
Inicio
clúster - SPSS Data E...
Estadística multivaria...
11. Análisis clúster - ...
11:57 a.m.


clúster - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
72 :
hoteles zocupaci zingreso ztrabaja zutilida zgastos zpuntoay zhabitac zgastene var
1 Abanico -1.39829 .81145 .79582 .43244 .55991 -.14647 -1.28964 -.11395
2 Estrella .85261 -1.99927 -1.64786 -.32332 -1.76432 -1.11317 1.01942 -1.26861
3 Río -.06181 1.63509 1.62368 1.15970 1.21701 -.14647 -.28702 1.68581
4 Varadero -.48386 -.07661 -.90977 .75503 -.20105 1.78693 -1.43395 -1.26759
5 Palma .57125 .96114 -.08191 .85535 .66556 .82023 -.94024 .82985
6 Playa -1.18726 .21814 .37691 -.13637 .08870 .82023 .83712 -.13283
7 Mariposa -.97624 1.37502 1.08508 -1.29680 1.11510 -.14647 1.24728 1.41653
8 Laguna .00853 -.91885 -.09188 1.18478 -.92862 -.14647 -1.16051 -.34072
9 Cayo 1.76704 .63846 -1.48828 .67980 .40293 -2.07987 -.90226 .93375
10 Arenas 1.13398 1.00452 -.10186 -.12953 .74867 -1.11317 -.18068 1.34111
11 Princesa -1.18726 -1.34772 -.83995 -1.20675 -.49102 -1.11317 -.94024 -1.33060
12 Vientos -.55420 .87574 .44672 -1.25006 .74840 -.14647 .06997 .09164
13 Piedra -1.39829 .42518 1.73340 .57607 .53084 .82023 1.21690 .47460
14 Orilla -.34318 -.60299 .08765 1.18706 -.24936 -1.11317 -.84150 1.30778
15 Coral .43057 .74870 -1.10926 -1.25234 .54067 -.14647 -1.06177 .56813
16 Caleta 1.20432 -2.70707 .08765 1.21784 .62243 .82023 1.32324 -.37660
17 Aguas .21955 .41546 .45670 -.67783 -.52820 .82023 .91308 -1.33077
18 Lago -.55420 -.66715 1.40425 .52135 -.34603 -.14647 -.43893 .00475
19 Luna -.34318 -.94588 -1.11923 1.28851 .46086 1.78693 .09276 .56813
20 Duna 1.27466 .46855 .08765 -1.20675 .63924 -.14647 1.30805 1.67714
21 Tocororo 1.76704 -.21884 .45670 .32073 .52338 -1.11317 -.59084 -.00699
Data View Variable View
SPSS Processor is ready
Inicio
clúster - SPSS ...
Output2 - SPS...
Estadística mul...
11. Análisis clú...
11:58 a.m.

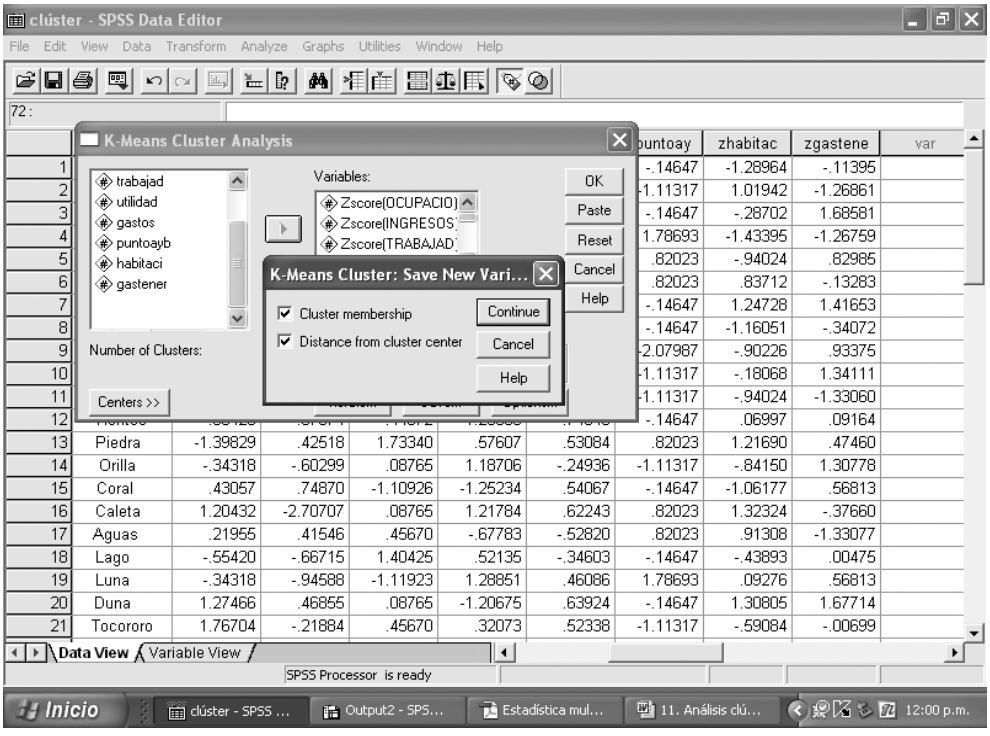
En la imagen anterior, se observa que a la base de datos original, el programa ha añadido ocho nuevas columnas que hacen referencia a las ocho variables de análisis pero ya estandarizadas.



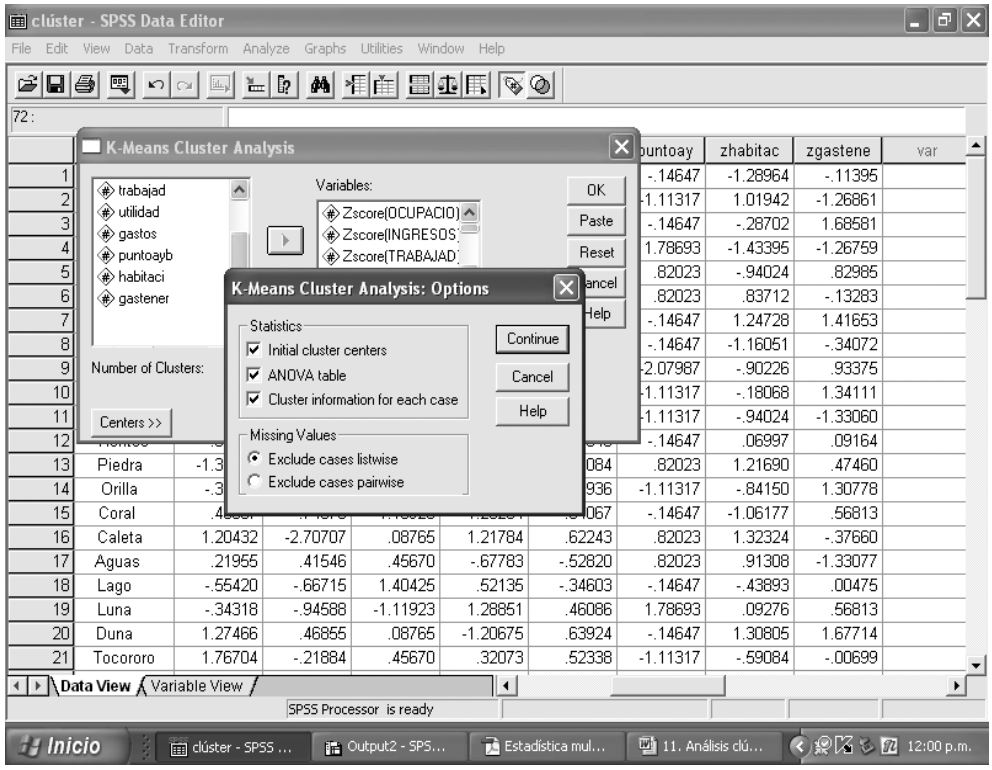


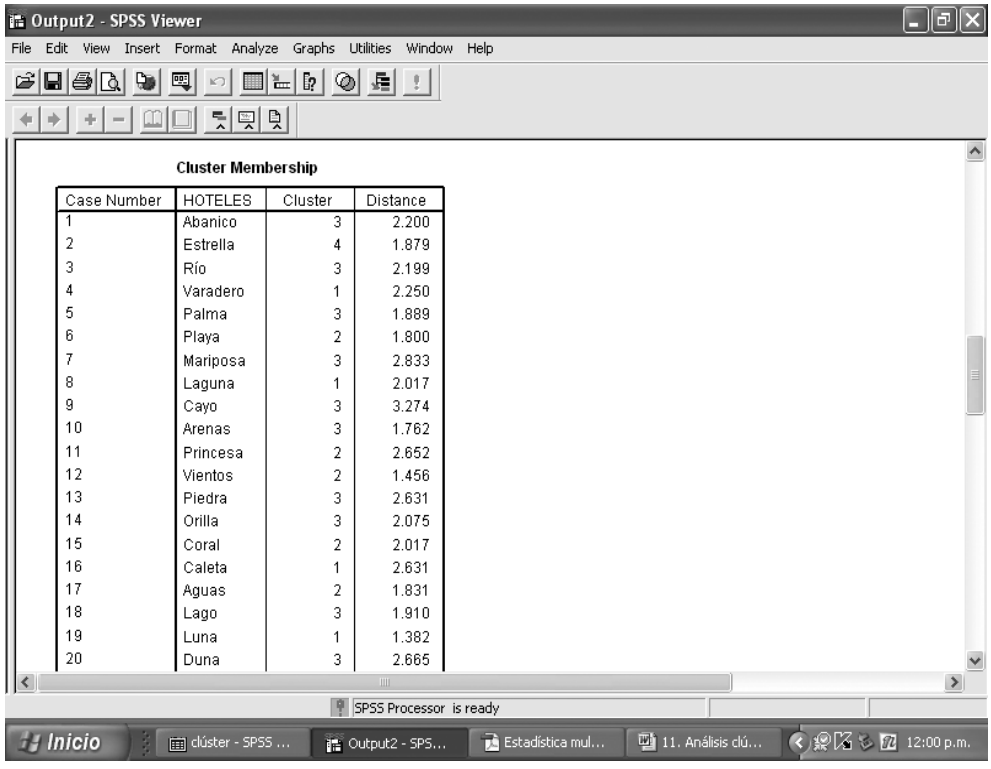


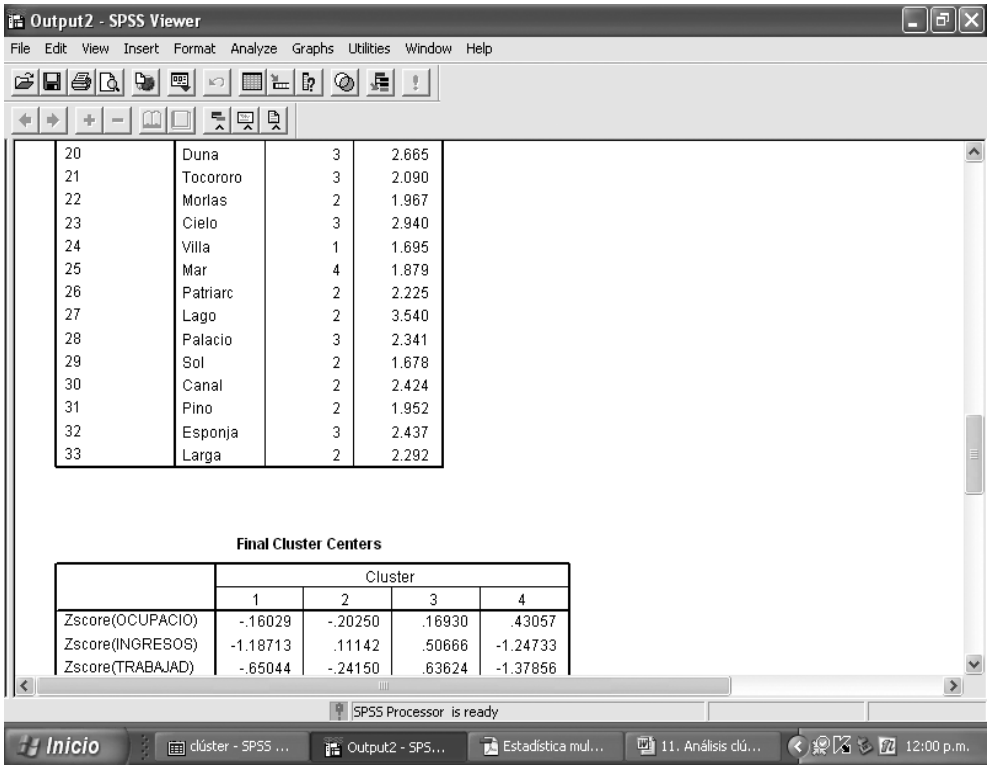
1							puntoay	zhabitac	zgastene	var
2							-.14647	-1.28964	-.11395	
3							-1.11317	1.01942	-1.26861	
4							-.14647	-.28702	1.68581	
5							1.78693	-1.43395	-1.26759	
6							.82023	-.94024	.82985	
7							.82023	.83712	-.13283	
8							-.14647	1.24728	1.41653	
9							-.14647	-1.16051	-.34072	
10							-2.07987	-.90226	.93375	
11							-1.11317	-.18068	1.34111	
12							-1.11317	-.94024	-1.33060	
13	Piedra	-1.39829	.42518	1.73340	.57607	.53084	-.14647	.06997	.09164	
14	Orilla	-.34318	-.60299	.08765	1.18706	-.24936	.82023	1.21690	.47460	
15	Coral	.43057	.74870	-1.10926	-1.25234	.54067	-1.11317	-.84150	1.30778	
16	Caleta	1.20432	-2.70707	.08765	1.21784	.62243	-.14647	-1.06177	.56813	
17	Aguas	.21955	.41546	.45670	-.67783	-.52820	.82023	1.32324	-.37660	
18	Lago	-.55420	-.66715	1.40425	.52135	-.34603	.82023	.91308	-1.33077	
19	Luna	-.34318	-.94588	-1.11923	1.28851	.46086	-.14647	-.43893	.00475	
20	Duna	1.27466	.46855	.08765	-1.20675	.63924	1.78693	.09276	.56813	
21	Tocororo	1.76704	-.21884	.45670	.32073	.52338	-.14647	1.30805	1.67714	
							-1.11317	-.59084	-.00699	



1							puntoay	zhabitac	zgastene	var
2							-.14647	-1.28964	-.11395	
3							-1.11317	1.01942	-1.26861	
4							-.14647	-.28702	1.68581	
5							1.78693	-1.43395	-1.26759	
6							.82023	-.94024	.82985	
7							.82023	.83712	-.13283	
8							-.14647	1.24728	1.41653	
9							-.14647	-1.16051	-.34072	
10							-2.07987	-.90226	.93375	
11							-1.11317	-.18068	1.34111	
12							-1.11317	-.94024	-1.33060	
13	Piedra	-1.39829	.42518	1.73340	.57607	.53084	-.14647	.06997	.09164	
14	Orilla	-.34318	-.60299	.08765	1.18706	-.24936	.82023	1.21690	.47460	
15	Coral	.43057	.74870	-1.10926	-1.25234	.54067	-1.11317	-.84150	1.30778	
16	Caleta	1.20432	-2.70707	.08765	1.21784	.62243	-.14647	-1.06177	.56813	
17	Aguas	.21955	.41546	.45670	-.67783	-.52820	.82023	1.32324	-.37660	
18	Lago	-.55420	-.66715	1.40425	.52135	-.34603	.82023	.91308	-1.33077	
19	Luna	-.34318	-.94588	-1.11923	1.28851	.46086	-.14647	-.43893	.00475	
20	Duna	1.27466	.46855	.08765	-1.20675	.63924	1.78693	.09276	.56813	
21	Tocororo	1.76704	-.21884	.45670	.32073	.52338	-.14647	1.30805	1.67714	
							-1.11317	-.59084	-.00699	

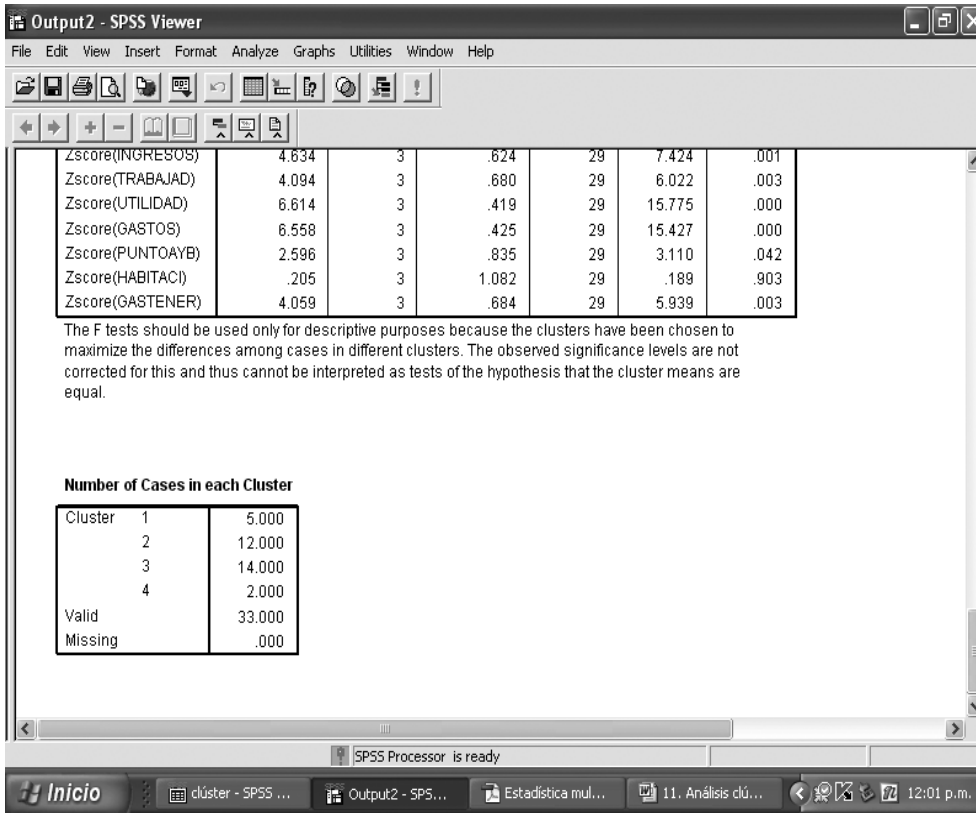

clúster - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
K-Means Cluster Analysis
trabajad
utilidad
gastos
puntoayb
habitaci
gastener
Variables:
Zscore(OCUPACIO)
Zscore(INGRESOS
Zscore(TRABAJAD
OK
Paste
Reset
Cancel
Help
Number of Clusters:
Centers >>
K-Means Cluster Analysis: Options
Statistics
Initial cluster centers
ANOVA table
Cluster information for each case
Continue
Cancel
Help
Missing Values
Exclude cases listwise
Exclude cases pairwise
puntoay zhabitac zgastene var
Piedra -1.3
Orilla -.3
Coral
Caleta 1.20432 -2.70707 .08765 1.21784 .62243 .82023 1.32324 -.37660
Aguas .21955 .41546 .45670 -.67783 -.52820 .82023 .91308 -1.33077
Lago -.55420 -.66715 1.40425 .52135 -.34603 -.14647 -.43893 .00475
Luna -.34318 -.94588 -1.11923 1.28851 .46086 1.78693 .09276 .56813
Duna 1.27466 .46855 .08765 -1.20675 .63924 -.14647 1.30805 1.67714
Tocororo 1.76704 -.21884 .45670 .32073 .52338 -1.11317 -.59084 -.00699
Data View Variable View
SPSS Processor is ready
Inicio
clúster - SPSS ...
Output2 - SPS...
Estadística mul...
11. Análisis clú...
12:00 p.m.


Output2 - SPSS Viewer
File Edit View Insert Format Analyze Graphs Utilities Window Help
Cluster Membership
Case Number HOTELES Cluster Distance
1 Abanico 3 2.200
2 Estrella 4 1.879
3 Río 3 2.199
4 Varadero 1 2.250
5 Palma 3 1.889
6 Playa 2 1.800
7 Mariposa 3 2.833
8 Laguna 1 2.017
9 Cayo 3 3.274
10 Arenas 3 1.762
11 Princesa 2 2.652
12 Vientos 2 1.456
13 Piedra 3 2.631
14 Orilla 3 2.075
15 Coral 2 2.017
16 Caleta 1 2.631
17 Aguas 2 1.831
18 Lago 3 1.910
19 Luna 1 1.382
20 Duna 3 2.665
SPSS Processor is ready
Inicio
clúster - SPSS ...
Output2 - SPS...
Estadística mul...
11. Análisis clú...
12:00 p.m.



En las dos imágenes anteriores, se muestra la tabla "*Cluster Membership*" donde aparece cada hotel asignado a su cluster. Véase que el:

- primer cluster está formado por los hoteles: Riu Varadero, Oasis Laguna Azul, Oasis Caleta Buena, Tryp Luna Plateada y Villa Real
- segundo cluster: Iberostar Playa Azul, Iberostar Princesa Roja, Meliá Fuertes Vientos, Sirenis Coral de Fuego, Meliá Aguas Claras, Iberostar Las Morlas, Paradisus Patriarca, Lago Verde, Barceló Sol Brillante, Oasis Canal Grande, Pino Alto y Playa Larga
- tercer cluster: Sirenis Abanico de Coral, Iberostar Río Azul, Tryp Palma Real, Paradisus Mariposa Blanca, Sol Varadero Azul, Sandals Arenas, Riu Piedra Dorada, Iberostar Orilla Azul, Sol Lago Azul, Meliá Duna Alta, Paradisus Tocororo, Tryp Cielo Azul, Sol Palacio y Sirenis Esponja de Mar
- cuarto cluster: Meliá Estrella de Mar y Riu Mar Profundo



En la imagen anterior, se muestra la tabla "*Number of Cases in each Cluster*" donde aparece un resumen final. Obsérvese que en el cluster 1 están aglomerados cinco hoteles, en el cluster 2 doce hoteles, en el cluster 3 catorce hoteles y en el cluster 4 sólo dos hoteles, para completar la muestra de treinta y tres instalaciones hoteleras en la que fueron estudiadas ocho variables.

EJERCITACIÓN

En el polo turístico de Varadero, la Delegación del Grupo Cubanacán, está realizando un estudio que incluye quince mercados emisores de turismo a Cuba. Basándose en los datos recopilados de ocho variables que han sido medidas en cada uno de los mercados, el objetivo de los miembros de la Delegación, consiste en agrupar dichos mercados según su similitud o semejanza en 5 clusters mediante el método jerárquico. Los datos son los siguientes:

Variables:

- gasto promedio mensual por habitante
- ingreso promedio mensual por habitante
- producto interno bruto per cápita
- % de la población que prefiere El Caribe como destino turístico
- % de la población que pertenece a la tercera edad y realiza viajes de turismo a Cuba anualmente
- % de la población que realiza viajes de turismo a Varadero anualmente
- % de la población que realiza viajes de turismo a Cuba anualmente (no incluye Varadero)
- % de la población que realiza viajes de turismo anualmente

Mercados	gasto	ingreso	pibpc	elcaribe	tercedad	varadero	cuba	turismo
Canadá	6972.45	7511.04	1077.25	27	15	69	23	59
Italia	1187.87	1988.39	1673.91	11	10	30	33	47
Australia	1091.75	1472.95	1349.17	18	9	22	35	43
Brasil	710.91	1517.98	2135.26	11	12	10	21	69
Alemania	1662.59	2017.92	1213.72	15	12	49	28	43
Francia	2549.79	3471.53	1361.50	17	10	28	28	44
México	1502.30	2226.11	1481.80	25	13	10	18	72
Colombia	527.59	752.87	1426.99	16	11	23	33	44
Rusia	6075.86	12183.92	2005.30	13	11	38	41	53
Argentina	1064.19	1097.44	993.66	26	13	32	20	48
Holanda	2730.85	3453.28	1264.54	16	10	24	20	36
Gran Bretaña	4966.85	10049.26	2023.27	13	12	67	27	72
Venezuela	1051.25	1326.15	1261.50	19	11	22	28	50
República Dominicana	519.90	939.12	1806.34	11	10	13	42	45
España	2106.78	3525.81	1673.56	19	10	36	47	47

SOLUCIÓN

- la distancia o diferencia mayor con un coeficiente igual a 52.218, es la existente entre los mercados Canadá y República Dominicana. Por el contrario, los más próximos o parecidos son los mercados Colombia y Venezuela con un coeficiente igual a 1.229
- según el gráfico de carámbanos, los diferentes clusters a cada nivel son:
  - Colombia y Venezuela
  - Francia y Holanda
  - Italia y República Dominicana
  - Colombia, Venezuela y Australia (primer multicluster)
  - etc.
- finalmente, el primer cluster está formado por el mercado: Canadá
- segundo cluster: Italia, Australia, Alemania, Francia, Colombia, Holanda, Venezuela, República Dominicana y España
- tercer cluster: Brasil
- cuarto cluster: México y Argentina
- quinto cluster: Rusia y Gran Bretaña

Capítulo
11

Gráficos de Pareto y Control.

11.1. Origen del Principio de Pareto.

A inicios del siglo XX, Vilfredo Pareto, un economista italiano, realizó un estudio sobre la riqueza y la pobreza descubriendo que el 20% de las personas, controlaba el 80% de la riqueza en Italia. Asimismo, observó muchas otras distribuciones similares en su estudio. El Principio, también conocido como la Regla 80-20, dice que el 20% de cualquier cosa, producirá el 80% de los efectos, mientras que el 80% restante, sólo cuenta para el 20% de los efectos.

Más tarde, comenzando la década de los años 50 de igual siglo, Joseph Juran descubrió la evidencia para la Regla de 80-20 en diversas situaciones. En particular, el fenómeno parecía existir sin excepción en problemas relacionados con la calidad.

11.2. ¿En qué consiste el Análisis de Pareto?

El Análisis de Pareto es una técnica que separa los "pocos vitales" de los "muchos triviales", de modo que el gráfico empleado, sirve para separar (visualmente) los aspectos significativos de un problema desde los triviales, para que un equipo sepa dónde dirigir sus esfuerzos para mejorar. Reducir los aspectos más significativos (las barras más largas en el gráfico de Pareto), servirá para una mejora general, que reducir los más pequeños.

11.3. ¿Cuándo utilizar un gráfico de Pareto?

Como el gráfico de Pareto es una herramienta de análisis de datos ampliamente utilizada, es útil en la determinación de la causa principal durante un esfuerzo de resolución de problemas.

Un equipo que esté llevando a cabo un proyecto para lograr mejoras continuas, puede emplearlo para los siguientes propósitos:

- analizar las causas
- estudiar los resultados
- planear una mejora continua
- demostrar qué progreso se ha logrado al servir como fotos del "antes y el después"

Véase un ejemplo.

Ejemplo 1:

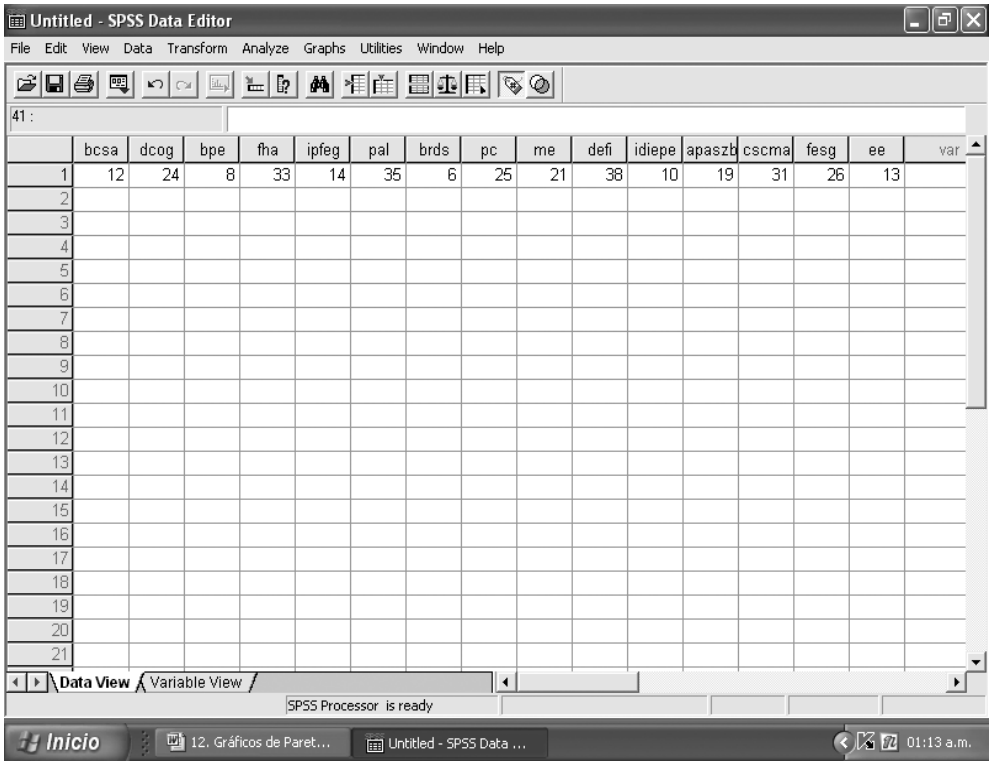
En el Hotel O, el equipo de trabajo que conforma el Comité de Calidad de la entidad, ha realizado una tormenta de ideas con numerosos trabajadores y especialistas de cada área, para determinar los principales problemas que están afectando a la entidad en estos momentos, y tratar de darles la más pronta solución por el negativo impacto económico que provocan, y también de imagen.

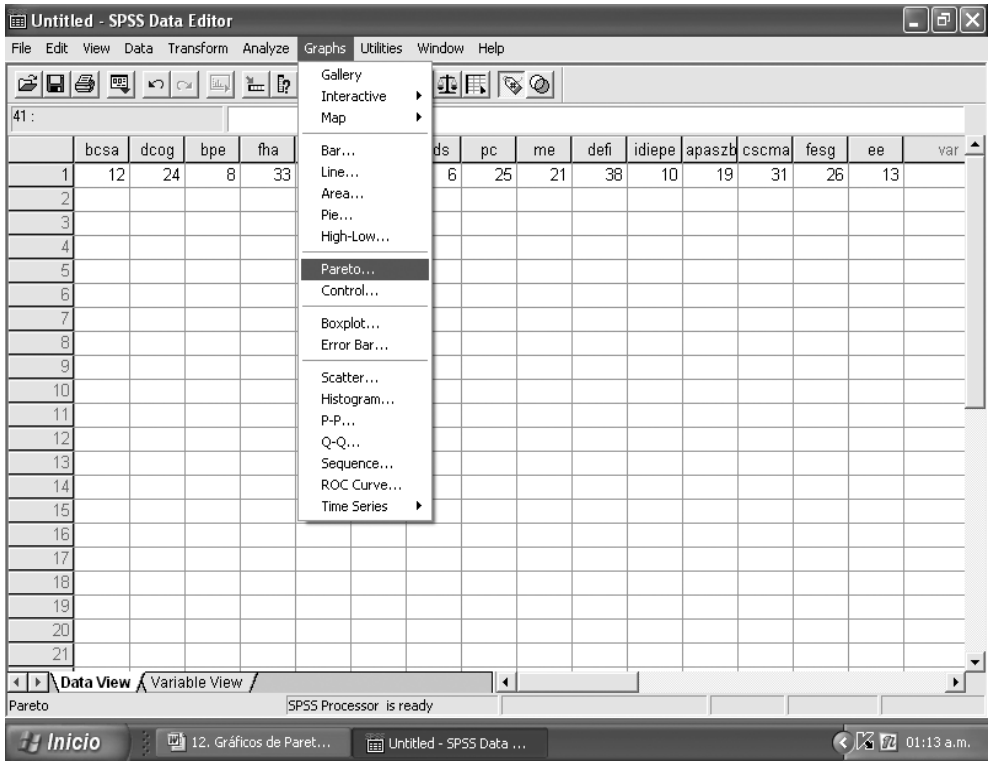
Los problemas mencionados fueron muchos, pero luego, quedaron agrupados en 15 categorías. A continuación, se observa la frecuencia con que fue mencionado cada problema (tarjado), lo cual servirá al comité para confeccionar el gráfico de Pareto.

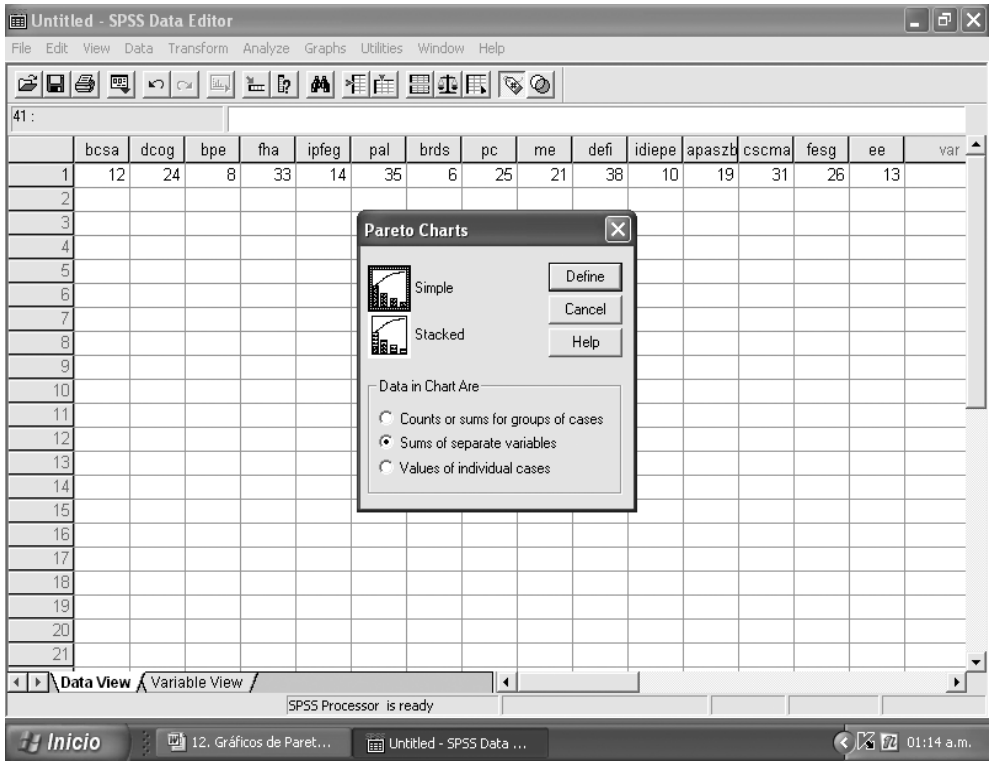
Problemas	Tarjado	n_i	f_i	N_i	F_i
Baja calidad del servicio de alojamiento	///// ///// //	12	0.038	12	0.038
Deficiente calidad de la oferta gastronómica	///// ///// ///// ///// ////	24	0.076	36	0.114
Baja profesionalidad de los empleados	///// ///	8	0.025	44	0.140
Falta de higienización de los alimentos	///// ///// ///// ///// ///// ///// ///	33	0.105	77	0.244
Inadecuada presencia física de los empleados de gastronomía	///// ///// ////	14	0.044	91	0.289
Pobre ambientación de los locales	///// ///// ///// ///// ///// ///// /////	35	0.111	126	0.400
Baja rapidez en los diferentes servicios	///// /	6	0.019	132	0.419
Poca comercialización	///// ///// ///// ///// /////	25	0.079	157	0.498
Mobiliario estropeado	///// ///// ///// ///// /	21	0.067	178	0.565
Deterioro del estado físico de la instalación	///// ///// ///// ///// ///// ///// ///// ///	38	0.121	216	0.686
Insuficiente dominio de idiomas extranjeros por parte de los empleados	///// /////	10	0.032	226	0.717
Ausencia de una piscina y área de servicio en la zona de bungalows	///// ///// ///// ////	19	0.060	245	0.778
Carencia de salón de conferencias con medios audiovisuales	///// ///// ///// ///// ///// ///// /	31	0.098	276	0.876
Falta de estabilidad en los suministros en general	///// ///// ///// ///// ///// /	26	0.083	302	0.959
Envejecimiento del equipamiento	///// ///// ///	13	0.041	315	1

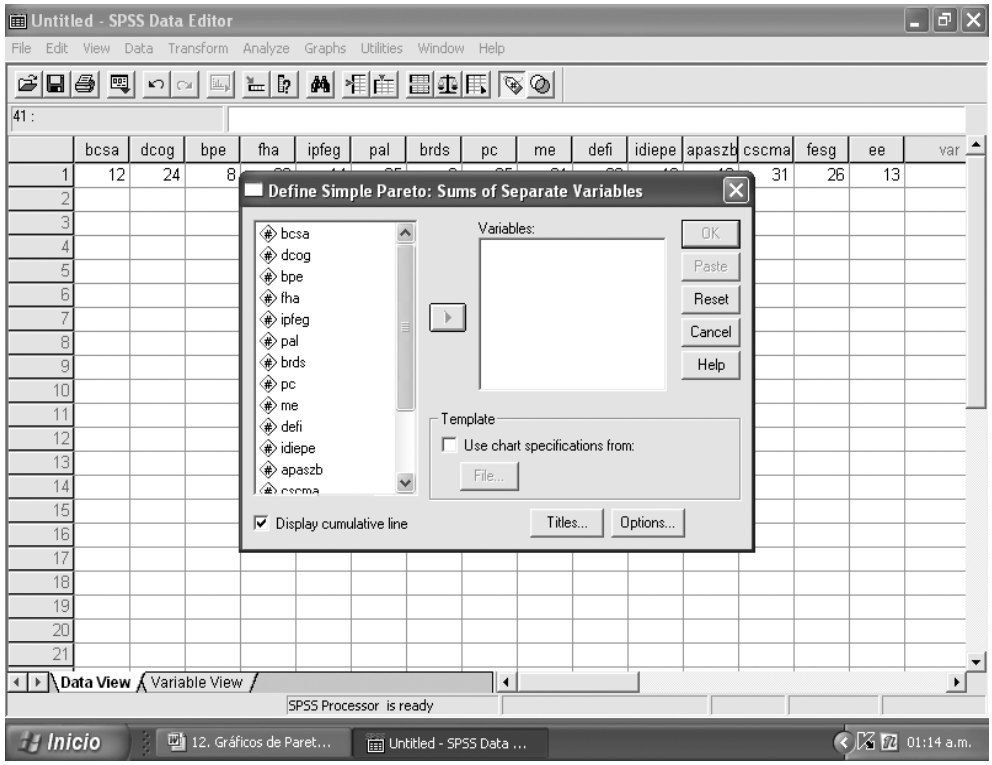
Solución:

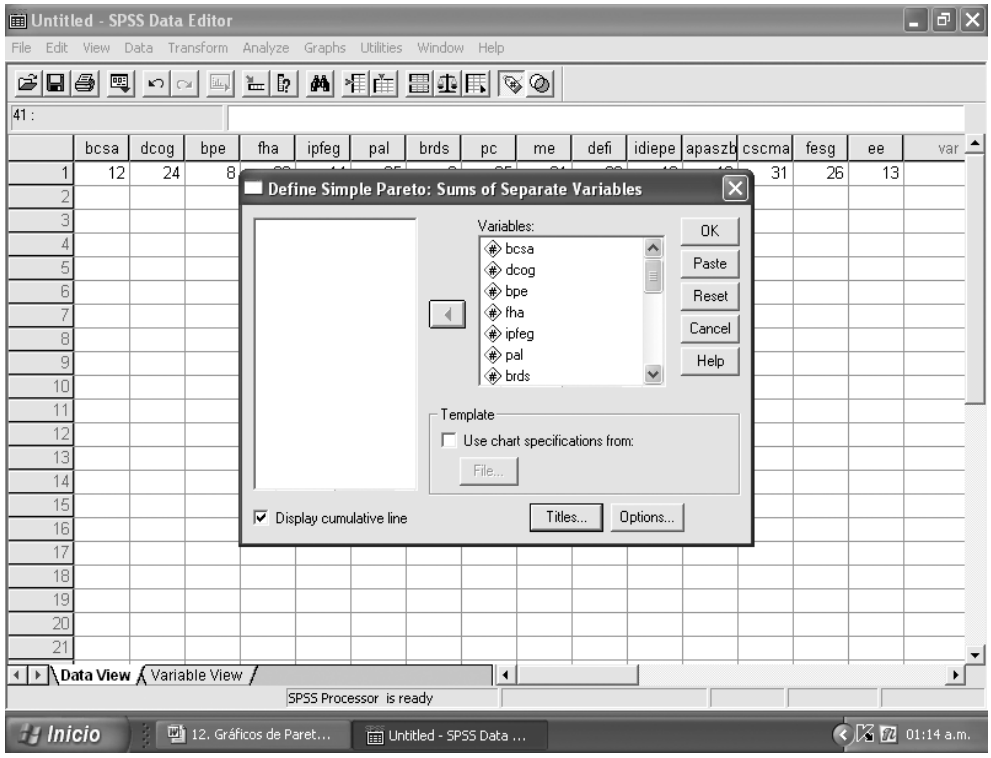
Empleando el SPSS, sería:

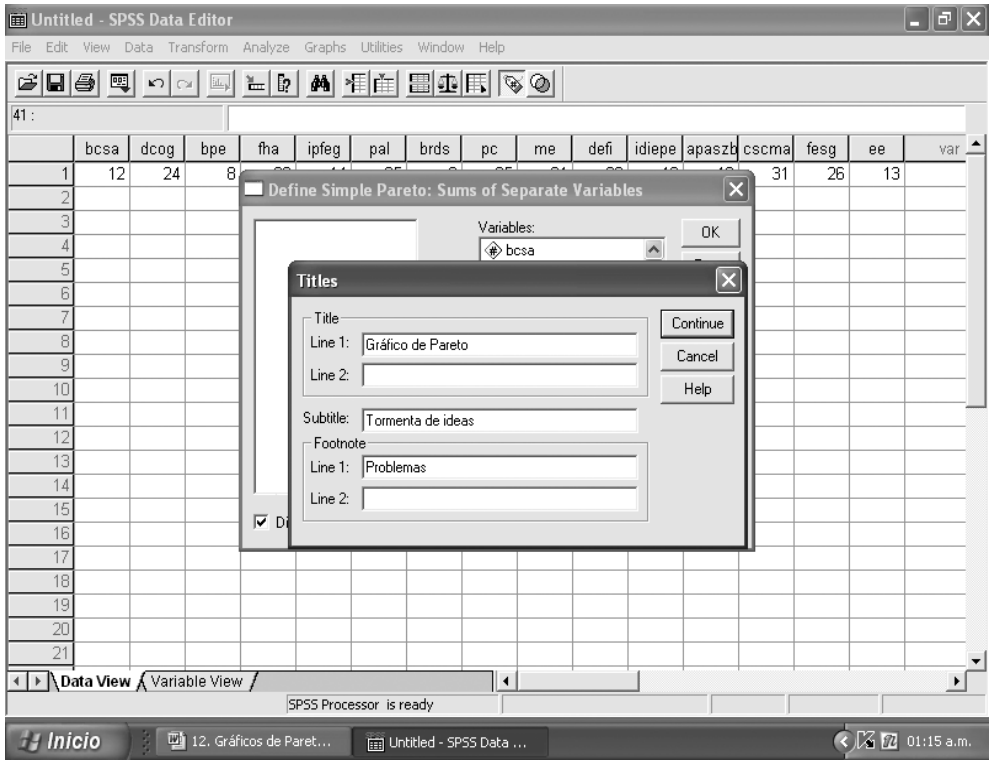

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
41 :
bcsa dcog bpe fha ipfeg pal brds pc me defi idiepe apaszb cscma fesg ee var
1 12 24 8 33 14 35 6 25 21 38 10 19 31 26 13
Data View Variable View
SPSS Processor is ready
Inicio
12. Gráficos de Paret...
Untitled - SPSS Data ...
01:13 a.m.

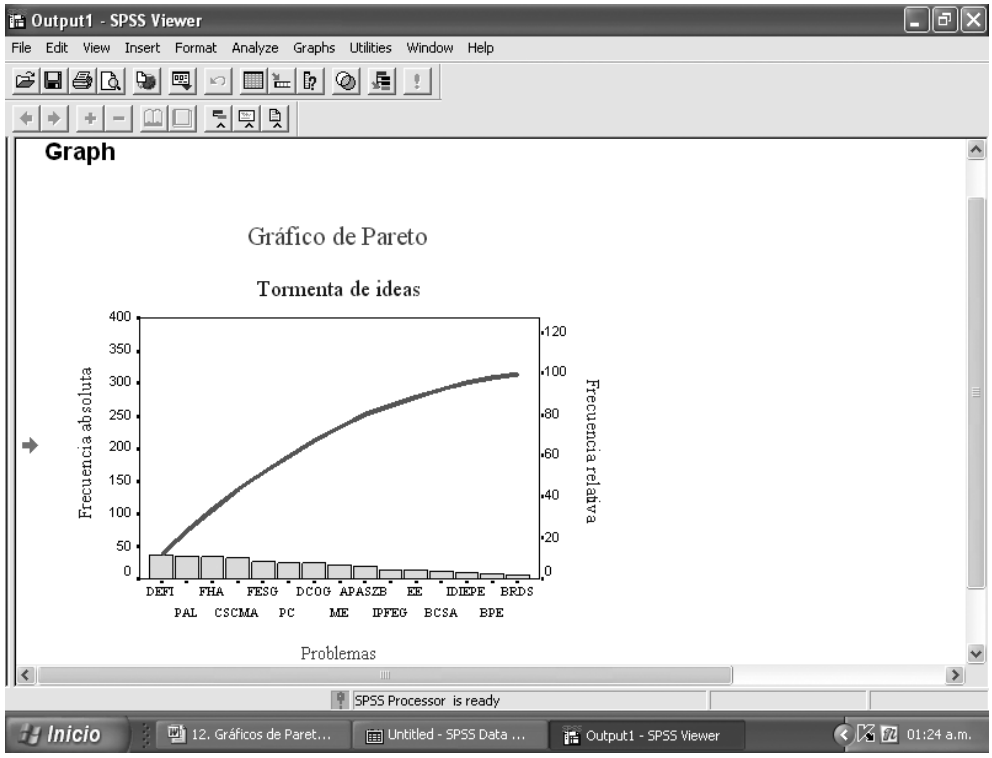

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Gallery
Interactive
Map
Bar...
Line...
Area...
Pie...
High-Low...
Pareto...
Control...
Boxplot...
Error Bar...
Scatter...
Histogram...
P-P...
Q-Q...
Sequence...
ROC Curve...
Time Series
41 :
bcsa dcog bpe fha ds pc me defi idiepe apaszb cscma fesg ee var
1 12 24 8 33 6 25 21 38 10 19 31 26 13
Data View Variable View
Pareto
SPSS Processor is ready
Inicio
12. Gráficos de Paret...
Untitled - SPSS Data ...
01:13 a.m.

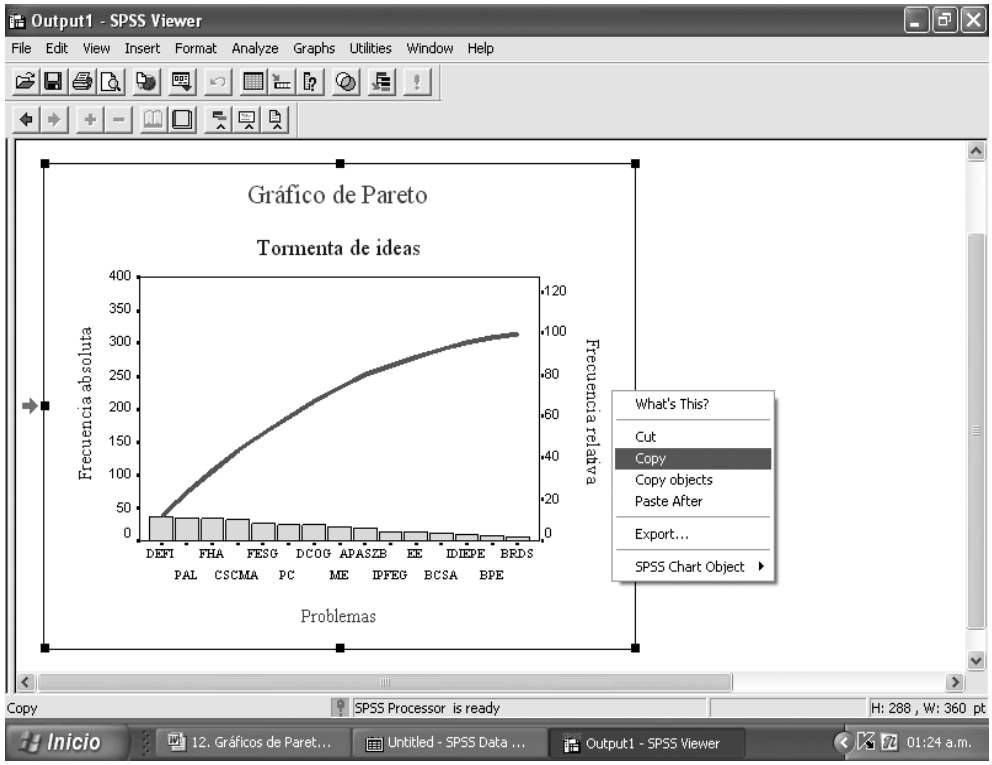

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
41 :
bcsa dcog bpe fha ipfeg pal brds pc me defi idiepe apaszb cscma fesg ee var
1 12 24 8 33 14 35 6 25 21 38 10 19 31 26 13
Pareto Charts
Simple
Stacked
Define
Cancel
Help
Data in Chart Are
Counts or sums for groups of cases
Sums of separate variables
Values of individual cases
Data View Variable View
SPSS Processor is ready
Inicio
12. Gráficos de Paret...
Untitled - SPSS Data ...
01:14 a.m.

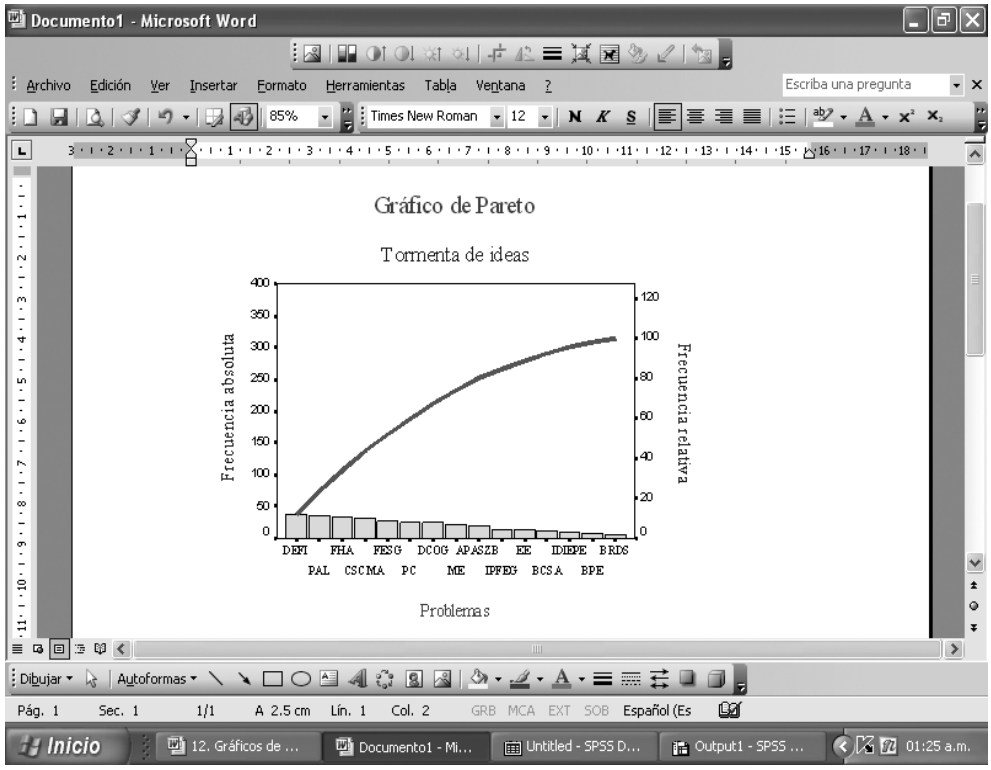

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
41 :
bcsa dcog bpe fha ipfeg pal brds pc me defi idiepe apaszb cscma fesg ee var
1 12 24 8 31 26 13
Define Simple Pareto: Sums of Separate Variables
bcsa
dcog
bpe
fha
ipfeg
pal
brds
pc
me
defi
idiepe
apaszb
Variables:
OK
Paste
Reset
Cancel
Help
Template
Use chart specifications from:
File...
Display cumulative line
Titles...
Options...
Data View Variable View
SPSS Processor is ready
Inicio
12. Gráficos de Paret...
Untitled - SPSS Data ...
01:14 a.m.

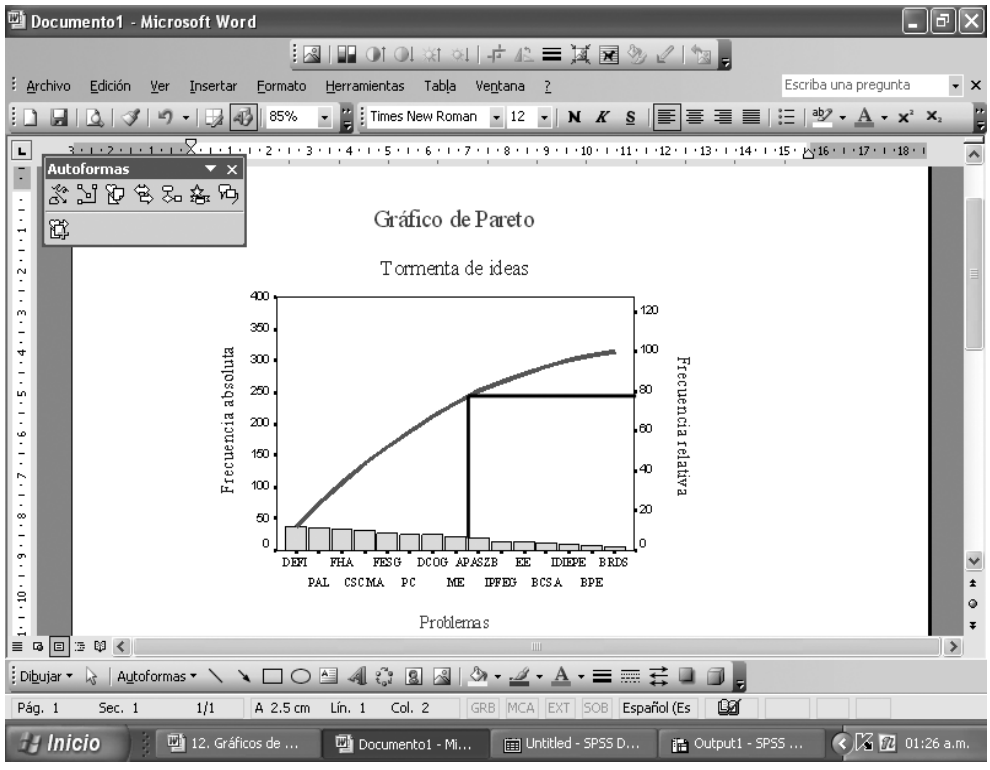

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Define Simple Pareto: Sums of Separate Variables
Variables:
bcsa
dcog
bpe
fha
ipfeg
pal
brds
OK
Paste
Reset
Cancel
Help
Template
Use chart specifications from:
File...
Display cumulative line
Titles...
Options...
Data View
Variable View
SPSS Processor is ready
Inicio
12. Gráficos de Paret...
Untitled - SPSS Data ...
01:14 a.m.


Untitled - SPSS Data Editor
Define Simple Pareto: Sums of Separate Variables
Titles
Title
Line 1: Gráfico de Pareto
Line 2:
Subtitle: Tormenta de ideas
Footnote
Line 1: Problemas
Line 2:
Continue
Cancel
Help
Data View
Variable View
SPSS Processor is ready
Inicio
12. Gráficos de Paret...
Untitled - SPSS Data ...
01:15 a.m.


Output1 - SPSS Viewer
File Edit View Insert Format Analyze Graphs Utilities Window Help
Graph
Gráfico de Pareto
Tormenta de ideas
Frecuencia absoluta
Frecuencia relativa
DEFI FHA FESG DCOG APASZB EE IDIEPE BRDS
PAL CSCMA PC ME IPFEG BCSA BPE
Problemas
SPSS Processor is ready
Inicio
12. Gráficos de Paret...
Untitled - SPSS Data ...
Output1 - SPSS Viewer
01:24 a.m.


Output1 - SPSS Viewer
File Edit View Insert Format Analyze Graphs Utilities Window Help
Gráfico de Pareto
Tormenta de ideas
Frecuencia absoluta
Frecuencia relativa
DEFI FHA FESG DCOG APASZB EE IDIEPE BRDS
PAL CSCMA PC ME IPFEG BCSA BPE
Problemas
What's This?
Cut
Copy
Copy objects
Paste After
Export...
SPSS Chart Object
Copy
SPSS Processor is ready
H: 288 , W: 360 pt
Inicio
12. Gráficos de Paret...
Untitled - SPSS Data ...
Output1 - SPSS Viewer
01:24 a.m.


Documento1 - Microsoft Word
Archivo Edición Ver Insertar Formato Herramientas Tabla Ventana ?
Escriba una pregunta
Times New Roman 12
Gráfico de Pareto
Tormenta de ideas
Frecuencia absoluta
Frecuencia relativa
DEFI FHA FESG DCOG APASZB EE IDIEPE BRDS
PAL CSCMA PC ME IPFEG BCSA BPE
Problemas
Dibujar Autoformas
Pág. 1 Sec. 1 1/1 A 2.5 cm Lín. 1 Col. 2 GRB MCA EXT SOB Español (Es
Inicio 12. Gráficos de ... Documento1 - Mi... Untitled - SPSS D... Output1 - SPSS ... 01:25 a.m.


Documento1 - Microsoft Word
Archivo Edición Ver Insertar Formato Herramientas Tabla Ventana ?
Escriba una pregunta
Times New Roman 12
Autoformas
Gráfico de Pareto
Tormenta de ideas
Frecuencia absoluta
Frecuencia relativa
DEFI FHA FESG DCOG APASZB EE IDIEPE BRDS
PAL CSCMA PC ME IPFEG BCSA BPE
Problemas
Dibujar Autoformas
Pág. 1 Sec. 1 1/1 A 2.5 cm Lín. 1 Col. 2 GRB MCA EXT SOB Español (Es
Inicio 12. Gráficos de ... Documento1 - Mi... Untitled - SPSS D... Output1 - SPSS ... 01:26 a.m.

problemas "pocos vitales" que requieren la más pronta solución por el negativo impacto económico y de imagen que causan al hotel, son los siguientes:

- deterioro del estado físico de la instalación
- pobre ambientación de los locales
- falta de higienización de los alimentos
- carencia de salón de conferencias con medios audiovisuales
- falta de estabilidad en los suministros en general
- poca comercialización
- deficiente calidad de la oferta gastronómica
- mobiliario estropeado

11.4. ¿Qué es un gráfico de control?

Un gráfico de control, conocido también como carta de control ideado por Walter Shewhart, es un gráfico que sirve para observar y analizar con datos estadísticos, la variabilidad y el comportamiento de un proceso a través del tiempo. Esto permitirá distinguir entre las variaciones por causas comunes y las especiales (atribuibles), lo que ayudará a caracterizar el funcionamiento del proceso, y así decidir las mejores acciones de control y mejora.

La variación por causas comunes (aleatorias), es aquella que permanece día a día inherente al proceso en sí, mientras que la variación por causas especiales, es causada por situaciones o circunstancias especiales que no son permanentes en el proceso. Por ello, un proceso que trabaja sólo con causas comunes de variación, se dice que está en control estadístico.

11.5. Límites de control.

Los límites no son las especificaciones o tolerancias para un proceso, por el contrario, se calculan a partir de la variación de los datos que se representa en el gráfico.

La clave está en establecer los límites para cubrir cierto porcentaje de la variación natural del proceso.

11.6. Tipos de gráficos de control.

Existen dos tipos generales de cartas de control:
- para variables
- para atributos

Los gráficos de control para variables, se aplican a variables continuas, y los más usuales son:
- X barra (de promedio)
- R (de rango)
- S (de desviación estándar)
- X (de medidas individuales)

Los gráficos de control para atributos, se aplican cuando el producto o proceso se juzga como conforme o no conforme, dependiendo de si posee ciertos atributos. Los más usuales son:
- p (proporción de artículos defectuosos)
- np (número de unidades defectuosas)
- c (número de defectos)
- u (número de defectos por unidad)

11.6.1. Gráfico de control para variable.

Véase un ejemplo.

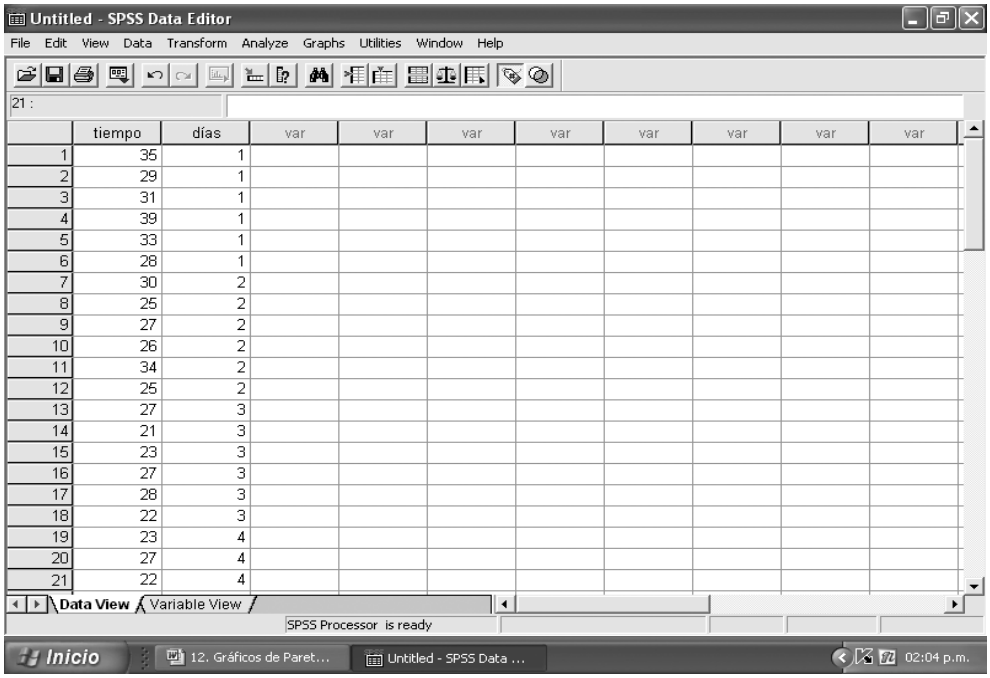
Ejemplo 2:

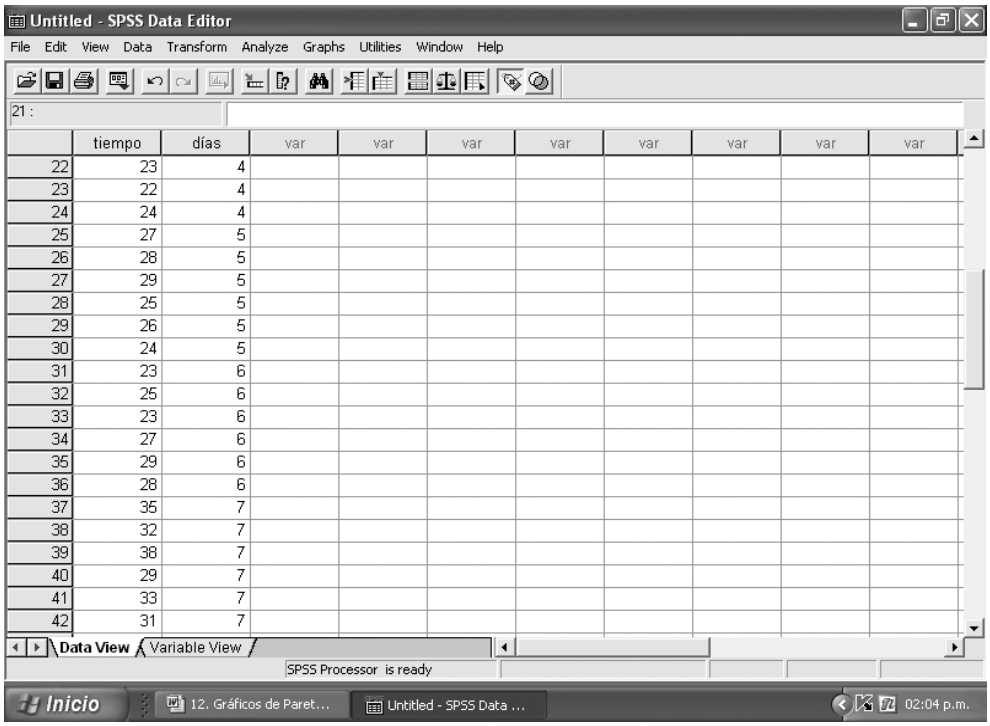
En un restaurante de la red de Palmares, se quiere analizar si durante los últimos siete días, el tiempo en que demoran los dependientes en servir los platos a los clientes, ha presentado alguna variación anormal con respecto a lo estandarizado. Para ello, se tomaron seis muestras de tiempos de demora diarios durante los siete días, como se muestra a continuación:

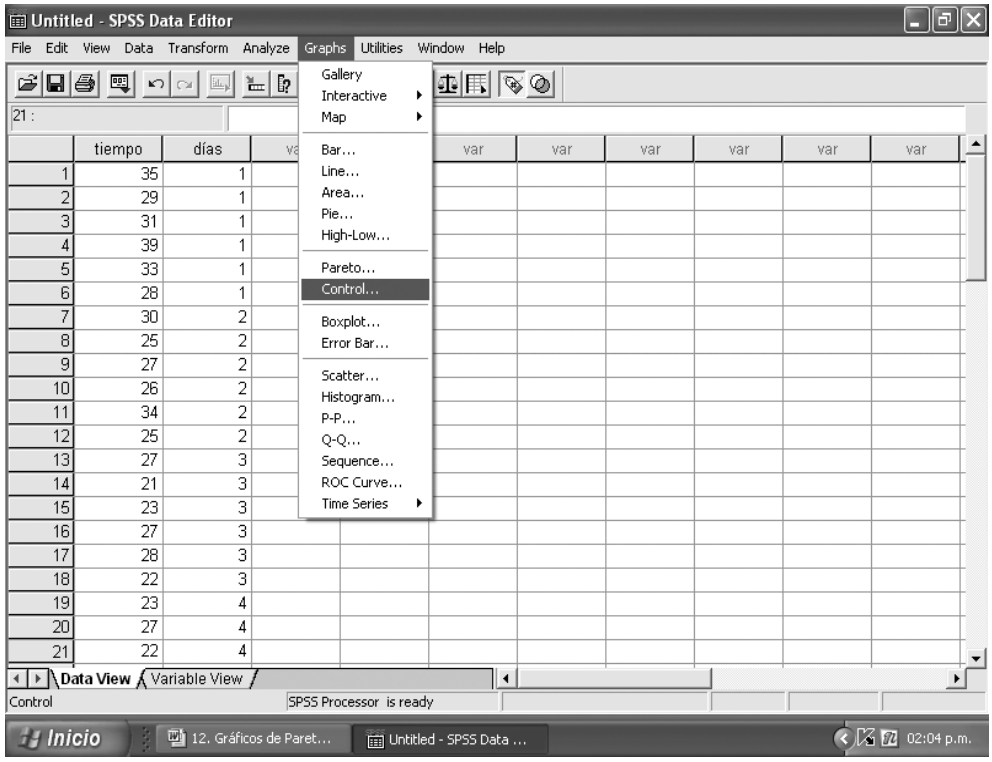
	Día 1	Día 2	Día 3	Día 4	Día 5	Día 6	Día 7
Tiempo (minutos)	35	30	27	23	27	23	35
	29	25	21	27	28	25	32
	31	27	23	22	29	23	38
	39	26	27	23	25	27	29
	33	34	28	22	26	29	33
	28	25	22	24	24	28	31

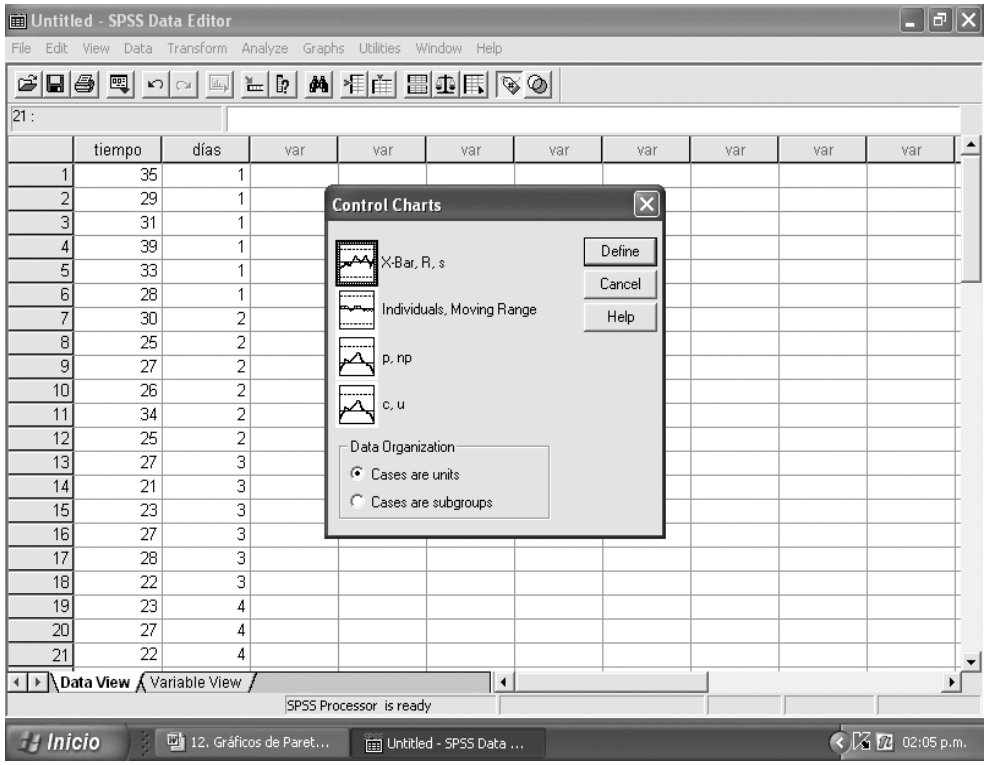
Solución:

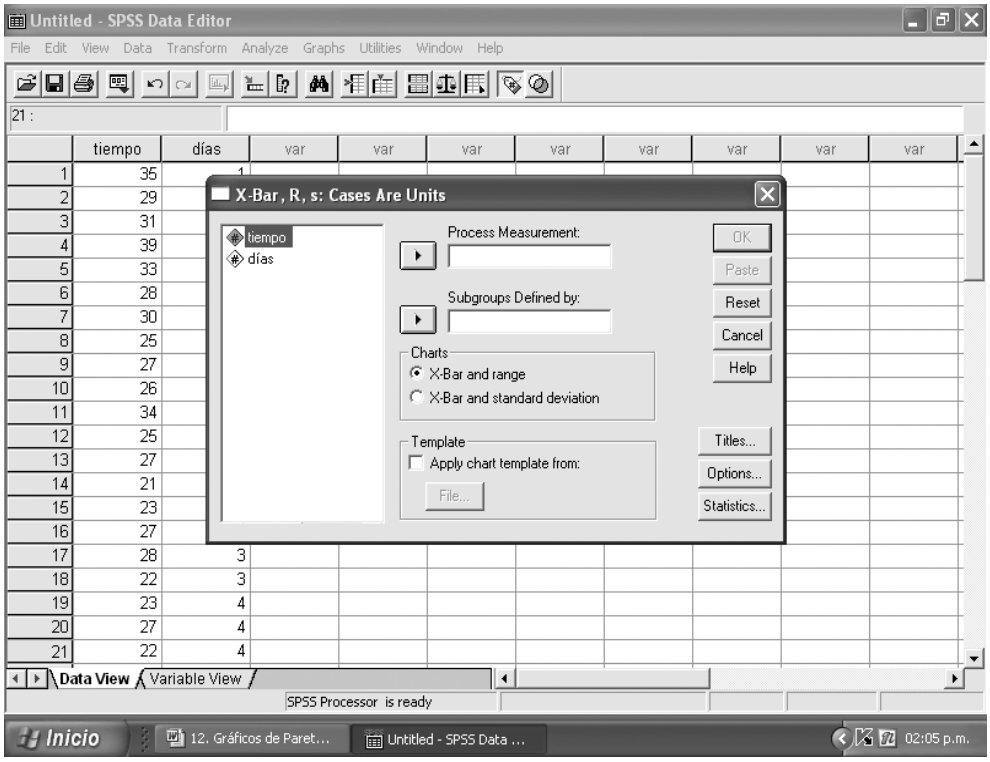
Utilizando el SPSS, sería:

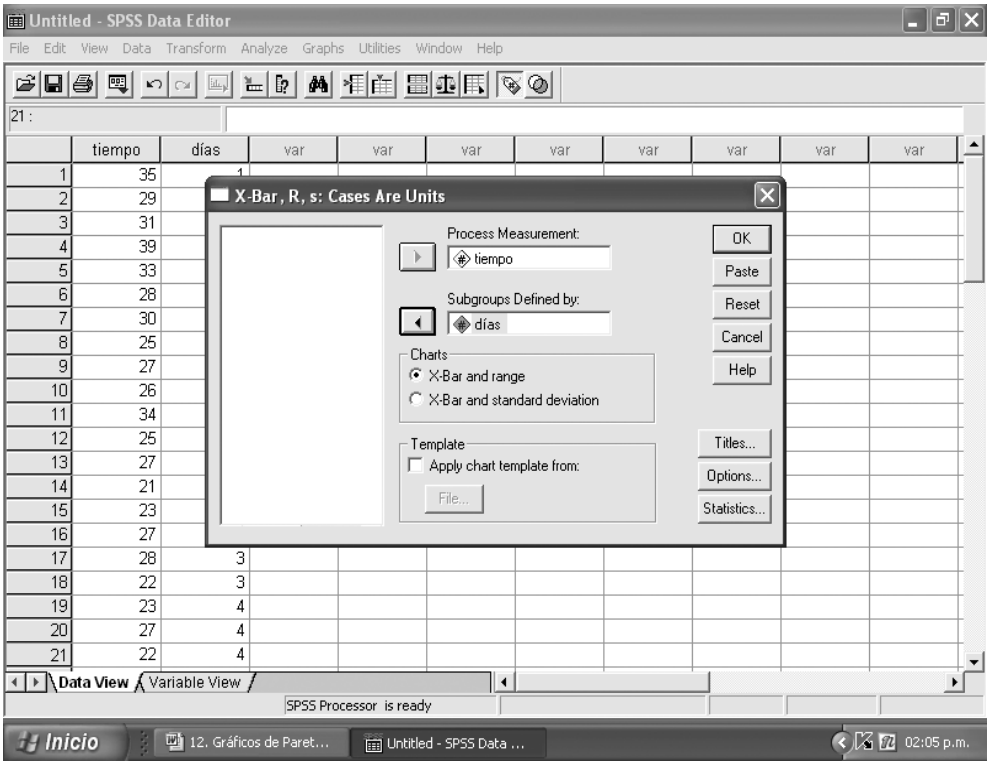


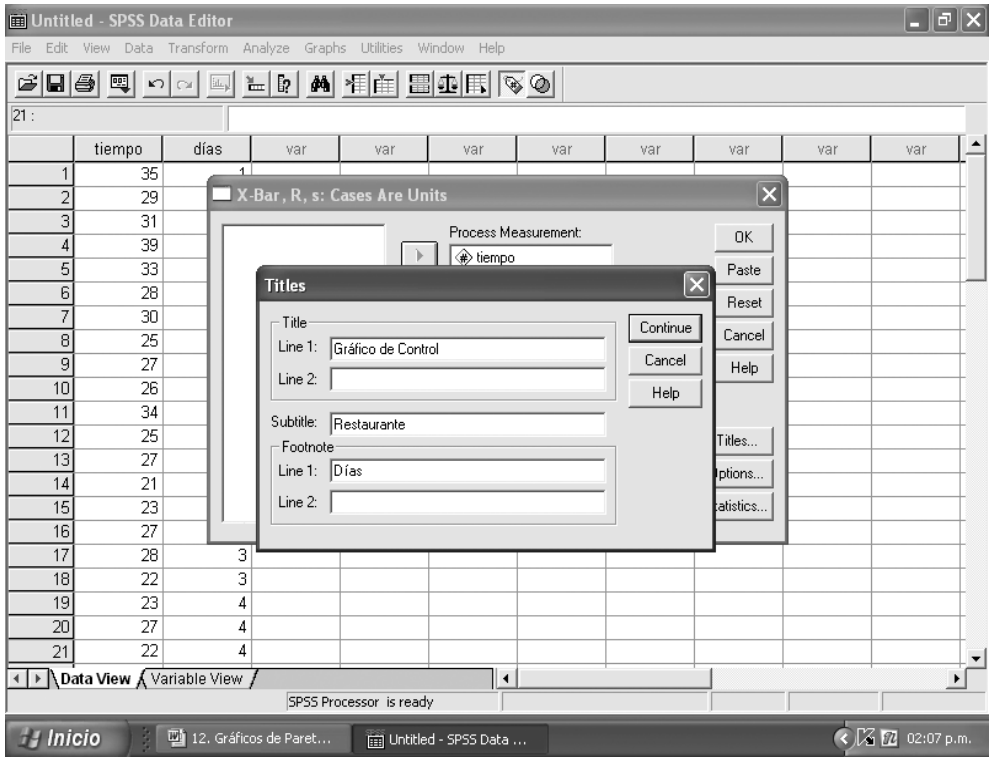


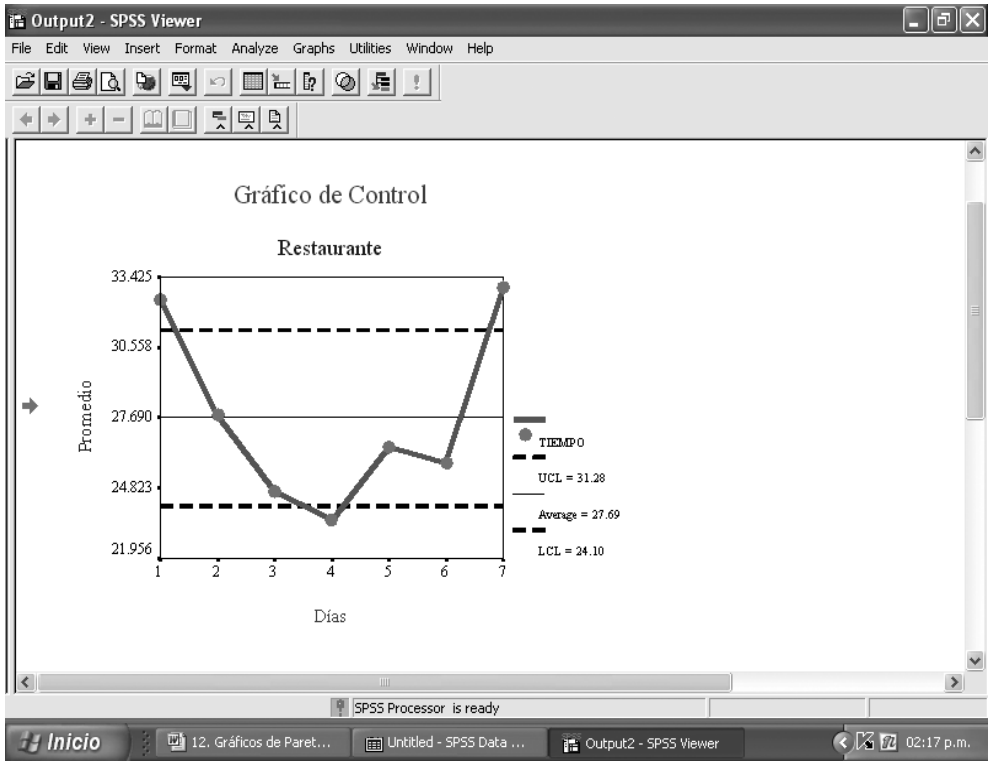

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Gallery
Interactive
Map
Bar...
Line...
Area...
Pie...
High-Low...
Pareto...
Control...
Boxplot...
Error Bar...
Scatter...
Histogram...
P-P...
Q-Q...
Sequence...
ROC Curve...
Time Series
21 :
tiempo
días
Data View
Variable View
Control
SPSS Processor is ready
Inicio
12. Gráficos de Paret...
Untitled - SPSS Data ...
02:04 p.m.

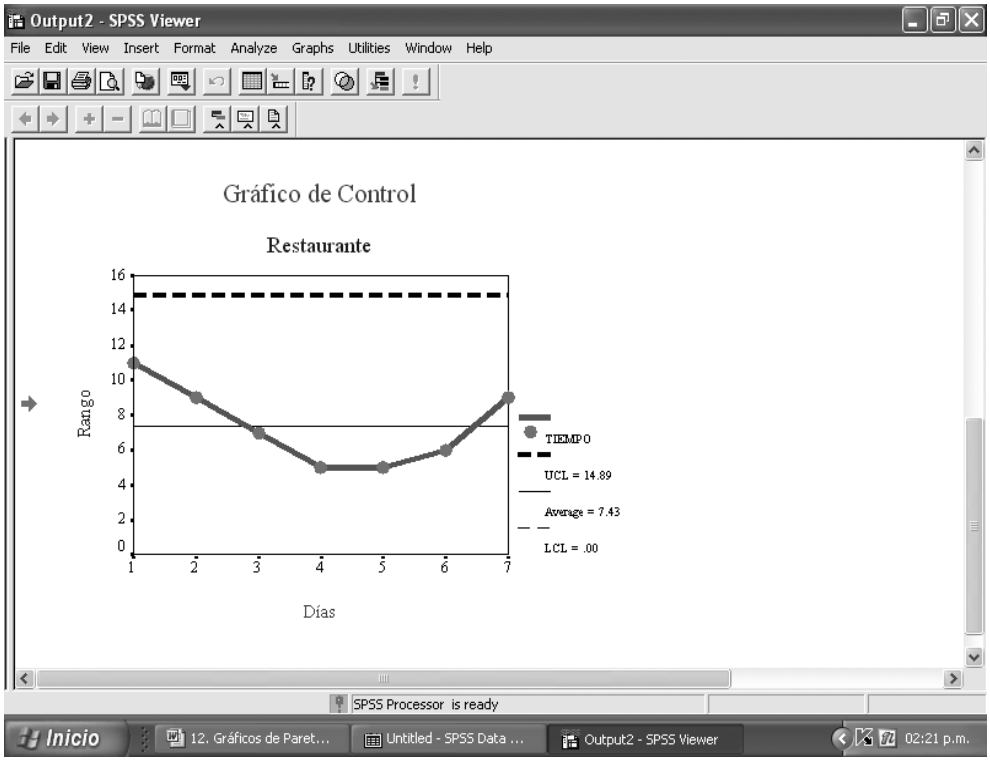

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
21 :
tiempo
días
Control Charts
X-Bar, R, s
Individuals, Moving Range
p, np
c, u
Define
Cancel
Help
Data Organization
Cases are units
Cases are subgroups
Data View
Variable View
SPSS Processor is ready
Inicio
12. Gráficos de Paret...
Untitled - SPSS Data ...
02:05 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
21 :
tiempo días var var var var var var var var
1 35 1
2 29
3 31
4 39
5 33
6 28
7 30
8 25
9 27
10 26
11 34
12 25
13 27
14 21
15 23
16 27
17 28 3
18 22 3
19 23 4
20 27 4
21 22 4
X-Bar, R, s: Cases Are Units
tiempo
días
Process Measurement:
Subgroups Defined by:
Charts
X-Bar and range
X-Bar and standard deviation
Template
Apply chart template from:
File...
OK
Paste
Reset
Cancel
Help
Titles...
Options...
Statistics...
Data View
Variable View
SPSS Processor is ready
Inicio
12. Gráficos de Paret...
Untitled - SPSS Data ...
02:05 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
21 :
tiempo días var var var var var var var var
1 35 1
2 29
3 31
4 39
5 33
6 28
7 30
8 25
9 27
10 26
11 34
12 25
13 27
14 21
15 23
16 27
17 28 3
18 22 3
19 23 4
20 27 4
21 22 4
X-Bar, R, s: Cases Are Units
Process Measurement:
tiempo
Subgroups Defined by:
días
Charts
X-Bar and range
X-Bar and standard deviation
Template
Apply chart template from:
File...
OK
Paste
Reset
Cancel
Help
Titles...
Options...
Statistics...
Data View
Variable View
SPSS Processor is ready
Inicio
12. Gráficos de Paret...
Untitled - SPSS Data ...
02:05 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
21 :
tiempo
días
var
X-Bar, R, s: Cases Are Units
Process Measurement:
tiempo
OK
Paste
Reset
Cancel
Help
Titles...
Options...
Statistics...
Titles
Title
Line 1: Gráfico de Control
Line 2:
Subtitle: Restaurante
Footnote
Line 1: Días
Line 2:
Continue
Cancel
Help
Data View
Variable View
SPSS Processor is ready
Inicio
12. Gráficos de Paret...
Untitled - SPSS Data ...
02:07 p.m.


Output2 - SPSS Viewer
File Edit View Insert Format Analyze Graphs Utilities Window Help
Gráfico de Control
Restaurante
33.425
30.558
27.690
24.823
21.956
Promedio
1
2
3
4
5
6
7
Días
TIEMPO
UCL = 31.28
Average = 27.69
LCL = 24.10
SPSS Processor is ready
Inicio
12. Gráficos de Paret...
Untitled - SPSS Data ...
Output2 - SPSS Viewer
02:17 p.m.



Véase en la penúltima imagen, el gráfico de control referido al promedio de la variable "tiempo". Se observan tres puntos fuera de los límites de control inferior y superior, correspondientes a los días 1 (30 minutos), 4 (23 minutos) y 7 (33 minutos). En la última imagen, se muestra el gráfico que representa el rango de la variable, y nótese que en ningún día hubo desviaciones significativas. En sentido general, los tiempos de demora del servicio que ofrecen los dependientes, no fluctúan en un rango muy amplio, pero sí vale la pena destacar que como hubo tres días alejados del promedio, entonces en esa semana no hubo estabilidad en el proceso de atención a los clientes por parte de los dependientes de servicio gastronómico. Los resultados hacen sospechar la presencia de causas especiales de variación en el proceso, por lo cual la dirección del restaurante, tendría que investigar cuáles son y qué las originan, para mejorarlo.

11.6.2. Gráfico de control para atributo.

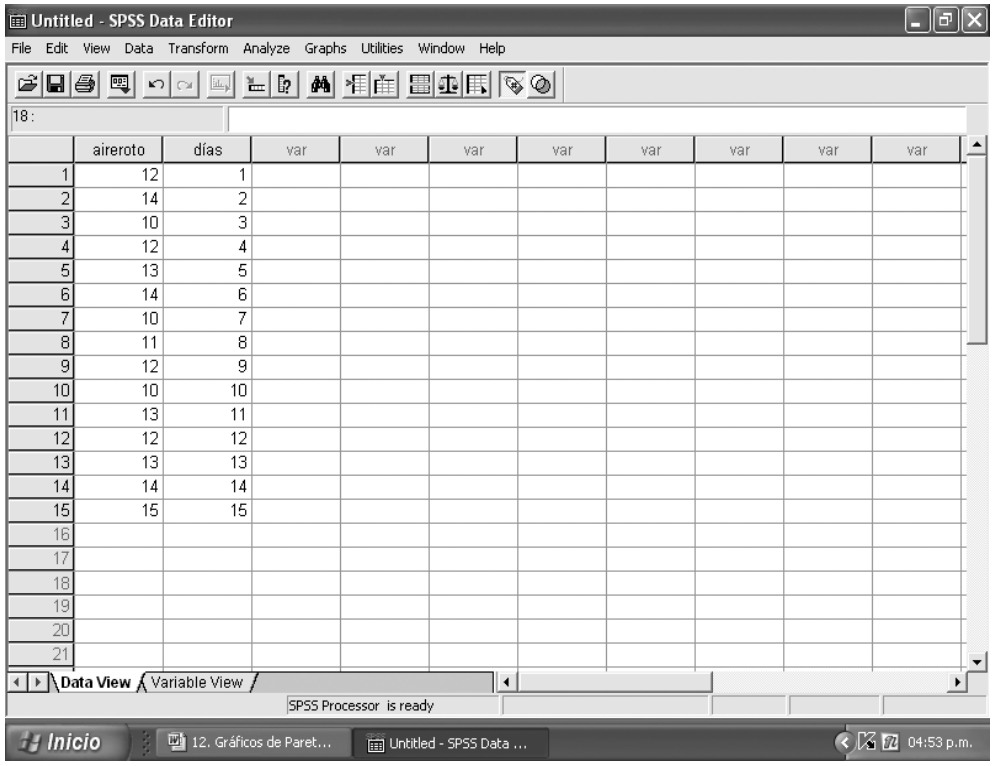
Véase un ejemplo.

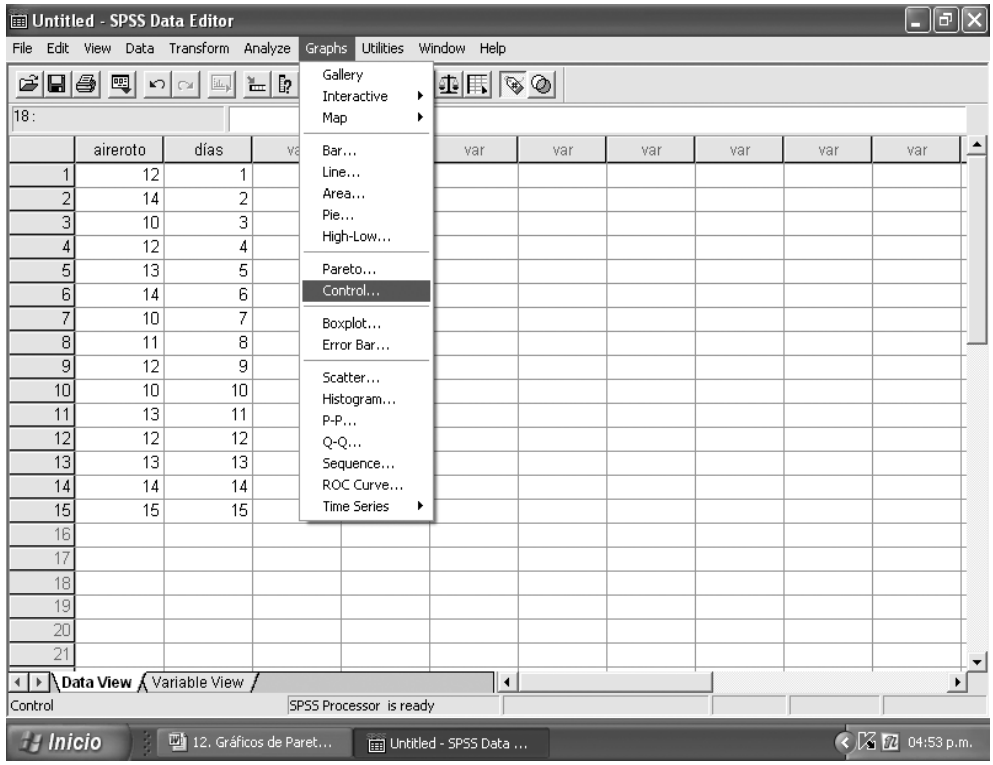
E*jemplo 3:*

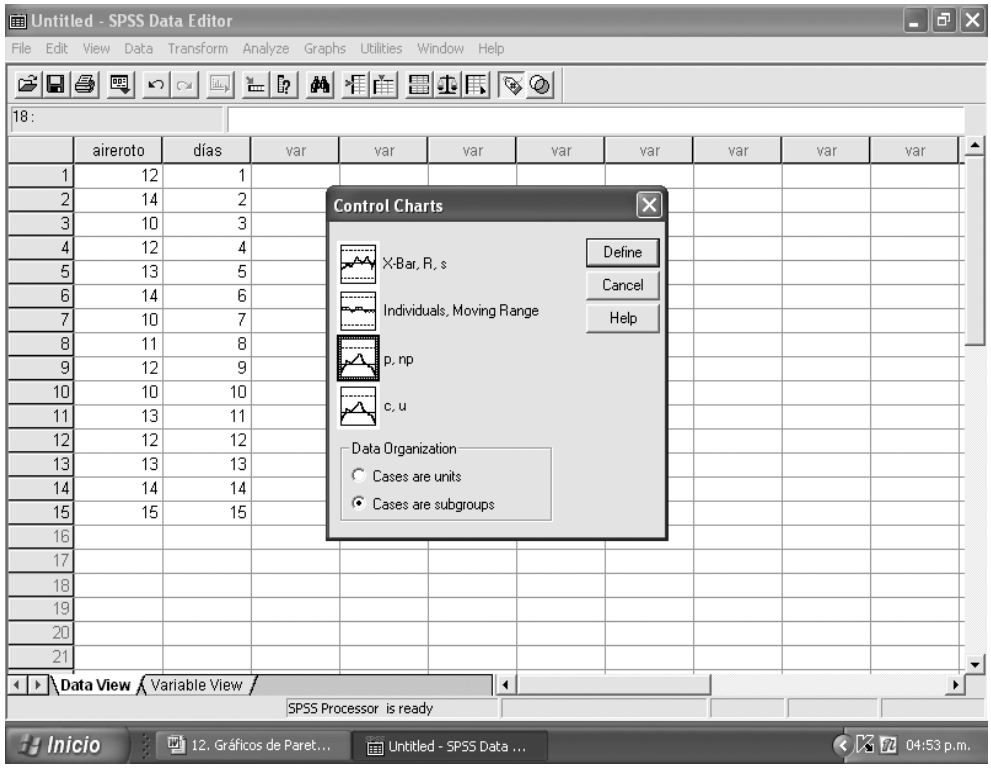
El Hotel P ubicado en el polo turístico de Cienfuegos, tiene un total de 402 habitaciones y se conoce que cada una, posee un equipo de aire acondicionado. Diariamente, el Departamento de Recepción recibe decenas de llamadas de clientes, informando de la rotura del aire acondicionado de su habitación, puesto que estos equipos deben ser cambiados por su mal estado técnico o recibir mantenimiento preventivo. A continuación, se muestran los datos correspondientes a quince días del mes pasado donde la recepcionista principal, recopiló la cantidad de roturas diarias, las cuales fueron luego, reportadas a Servicios Técnicos para su arreglo. Ella desea saber si hubo algún día que presentó demasiadas roturas como para desestabilizar el proceso de alojamiento en el hotel.

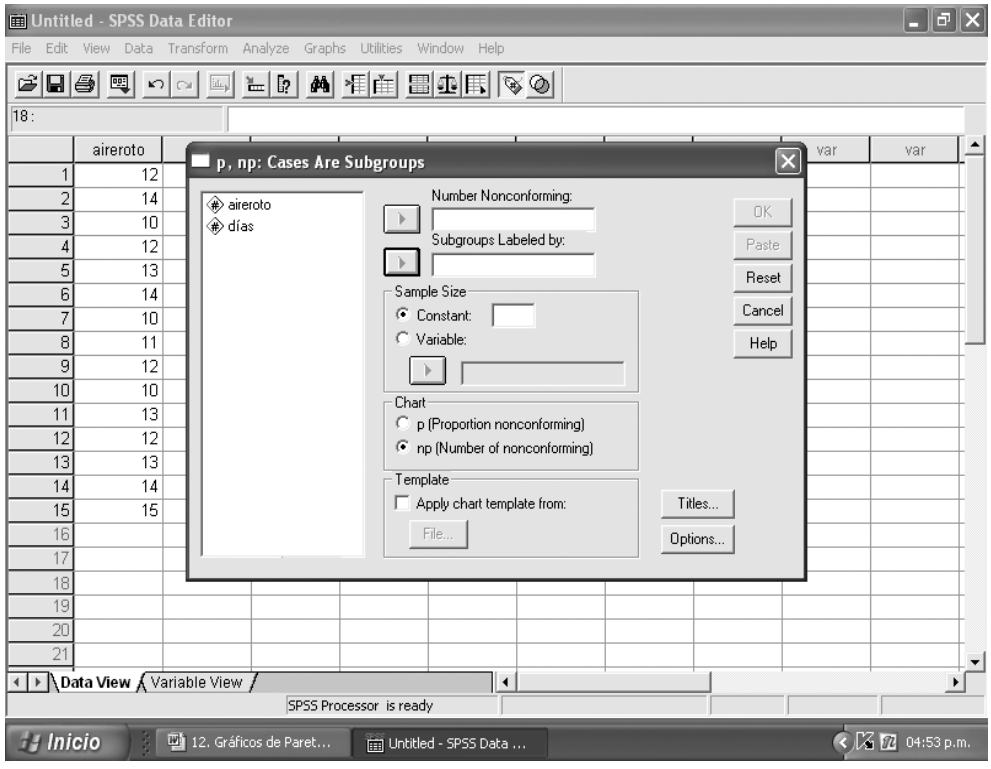
Días	Cantidad de equipos de aire acondicionado rotos
1	12
2	14
3	10
4	12
5	13
6	14
7	10
8	11
9	12
10	10
11	13
12	12
13	13
14	14
15	15

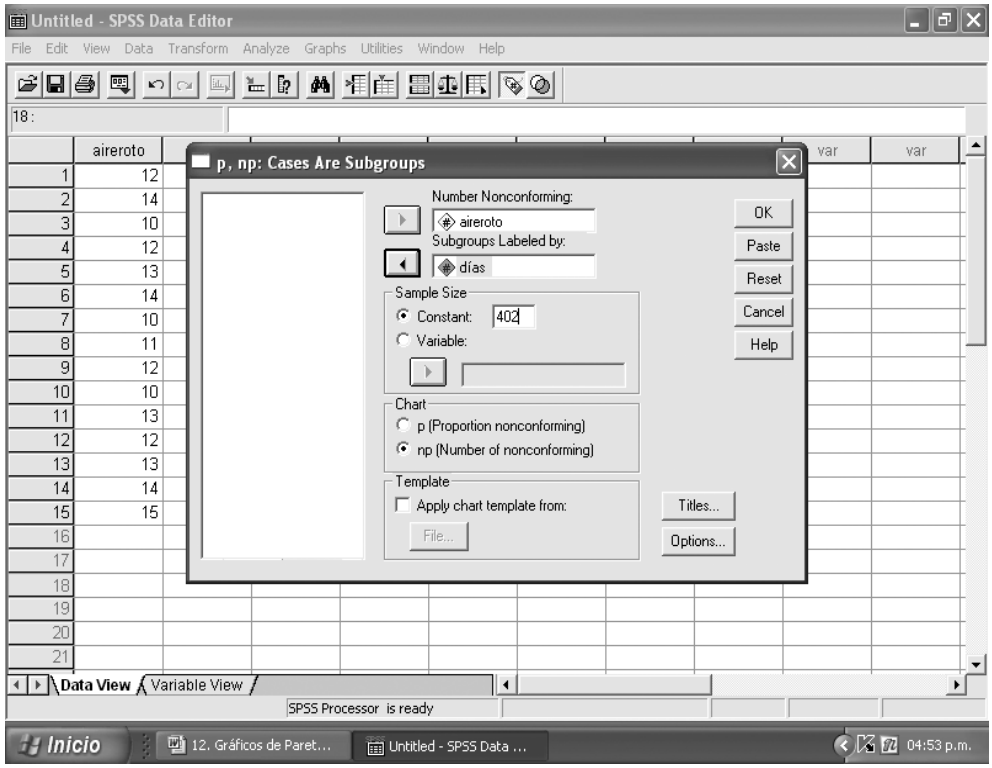
Solución:
Empleando el SPSS, sería:

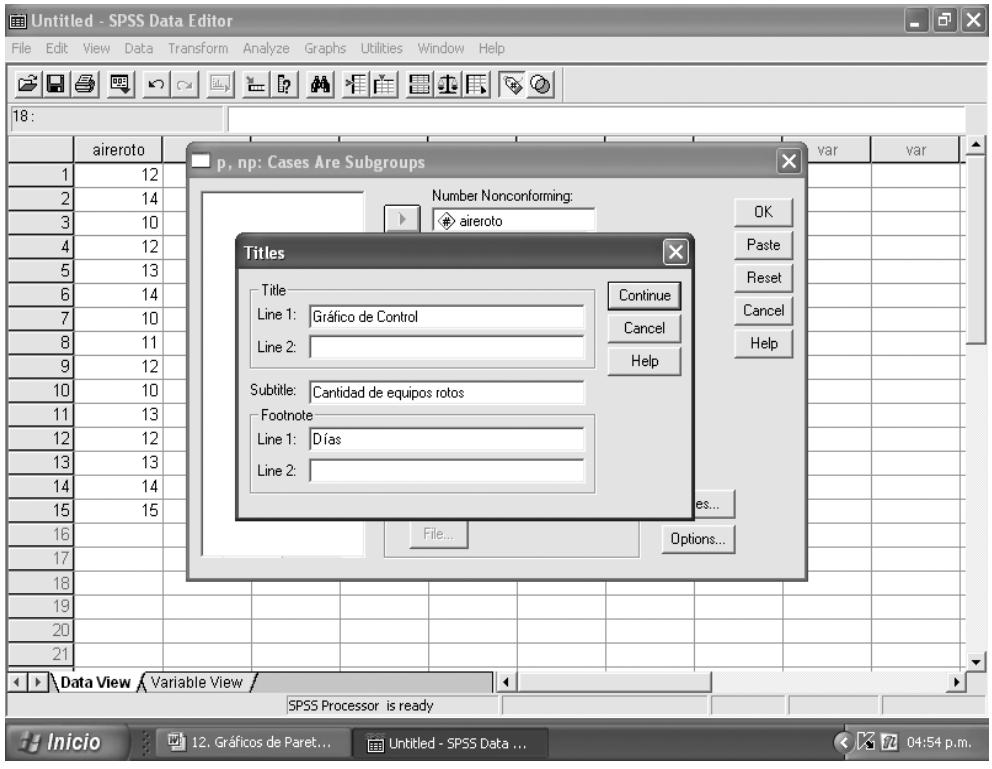

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
18 :
aireroto días var var var var var var var var
1 12 1
2 14 2
3 10 3
4 12 4
5 13 5
6 14 6
7 10 7
8 11 8
9 12 9
10 10 10
11 13 11
12 12 12
13 13 13
14 14 14
15 15 15
16
17
18
19
20
21
Data View Variable View
SPSS Processor is ready
Inicio
12. Gráficos de Paret...
Untitled - SPSS Data ...
04:53 p.m.

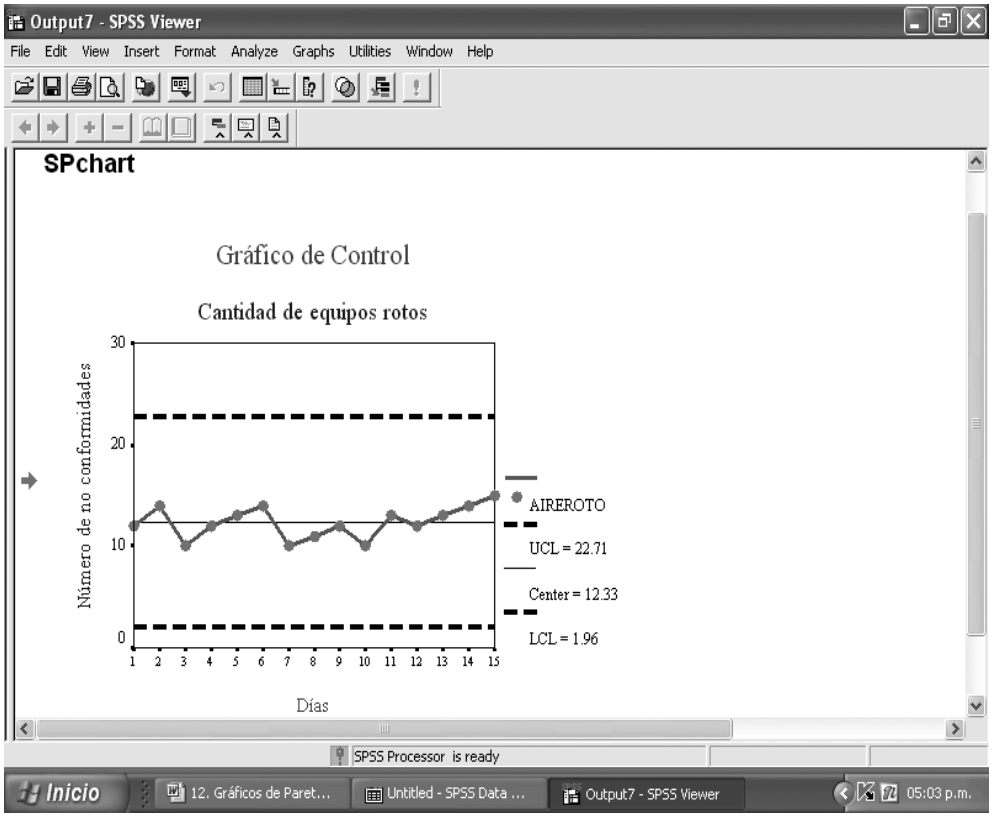

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Gallery
Interactive
Map
Bar...
Line...
Area...
Pie...
High-Low...
Pareto...
Control...
Boxplot...
Error Bar...
Scatter...
Histogram...
P-P...
Q-Q...
Sequence...
ROC Curve...
Time Series
18 :
aireroto días var var var var var var var
1 12 1
2 14 2
3 10 3
4 12 4
5 13 5
6 14 6
7 10 7
8 11 8
9 12 9
10 10 10
11 13 11
12 12 12
13 13 13
14 14 14
15 15 15
16
17
18
19
20
21
Data View Variable View
Control
SPSS Processor is ready
Inicio
12. Gráficos de Paret...
Untitled - SPSS Data ...
04:53 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
18 :
aireroto
días
var
Control Charts
X-Bar, R, s
Individuals, Moving Range
p, np
c, u
Define
Cancel
Help
Data Organization
Cases are units
Cases are subgroups
Data View
Variable View
SPSS Processor is ready
Inicio
12. Gráficos de Paret...
Untitled - SPSS Data ...
04:53 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
18 :
aireroto
var
p, np: Cases Are Subgroups
aireroto
días
Number Nonconforming:
Subgroups Labeled by:
Sample Size
Constant:
Variable:
Chart
p (Proportion nonconforming)
np (Number of nonconforming)
Template
Apply chart template from:
File...
OK
Paste
Reset
Cancel
Help
Titles...
Options...
Data View
Variable View
SPSS Processor is ready
Inicio
12. Gráficos de Paret...
Untitled - SPSS Data ...
04:53 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
18 :
aireroto
var
var
p, np: Cases Are Subgroups
Number Nonconforming:
aireroto
Subgroups Labeled by:
días
Sample Size
Constant: 402
Variable:
Chart
p (Proportion nonconforming)
np (Number of nonconforming)
Template
Apply chart template from:
File...
OK
Paste
Reset
Cancel
Help
Titles...
Options...
Data View
Variable View
SPSS Processor is ready
Inicio
12. Gráficos de Paret...
Untitled - SPSS Data ...
04:53 p.m.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
18 :
aireroto
var
var
p, np: Cases Are Subgroups
Number Nonconforming:
aireroto
Titles
Title
Line 1: Gráfico de Control
Line 2:
Subtitle: Cantidad de equipos rotos
Footnote
Line 1: Días
Line 2:
Continue
Cancel
Help
OK
Paste
Reset
Cancel
Help
File...
Options...
Data View
Variable View
SPSS Processor is ready
Inicio
12. Gráficos de Paret...
Untitled - SPSS Data ...
04:54 p.m.



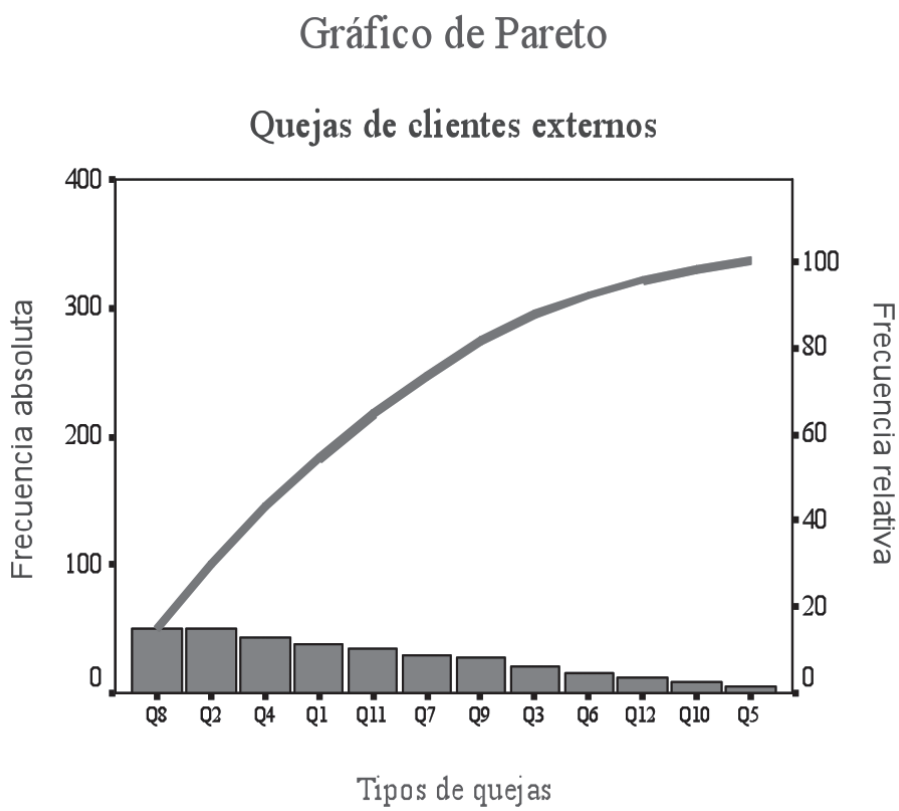
Según se observa en la imagen anterior, la cantidad de equipos de aire acondicionado rotos diariamente durante quince días, no crea descontrol en el proceso de alojamiento, puesto que no existe ningún punto fuera de los límites.

EJERCITACIÓN

El Departamento de Calidad y Atención al Cliente del Hotel W, ha recopilado las quejas expresadas por los clientes externos respecto a su estancia en la instalación. En sentido general, la mayoría de las quejas han estado relacionadas con los servicios de alojamiento, animación y alimentos y bebidas. La asistente del departamento ha agrupado las quejas en 12 tipos, y cada uno, con la cantidad respectiva de clientes que lo expresó. Ahora procederá a determinar, cuáles son los tipos de quejas de mayor impacto negativo o que provocan elevada insatisfacción en los clientes. Los datos se hallan tabulados a continuación:

	Tipos de quejas	Frecuencia absoluta
1	Animación nocturna poco divertida	38
2	Idioma deficiente de las recepcionistas	50
3	Baja profesionalidad de los porteros-maleteros	21
4	Demora en el servicio de habitación	44
5	Los elevadores son lentos	6
6	Los dependientes presentan demoras en el servicio	15
7	Horario demasiado restringido del buffet	30
8	Poca variedad de la comida	51
9	Habitaciones poco modernas en su interior	27
10	En el bar de la playa, la oferta de menú es poca	8
11	Servicio de check-in muy lento	34
12	No existe ningún punto de consumo abierto las 24 horas	12

SOLUCIÓN



Los tipos de quejas que provocan la mayor insatisfacción de los clientes (pocos vitales) elevando la posibilidad de que no repitan su estancia en la instalación ni recomienden a otros su visita, son los siguientes:

- poca variedad de la comida
- idioma deficiente de las recepcionistas
- demora en el servicio de habitación
- animación nocturna poco divertida
- servicio de check-in muy lento
- horario demasiado restringido del buffet

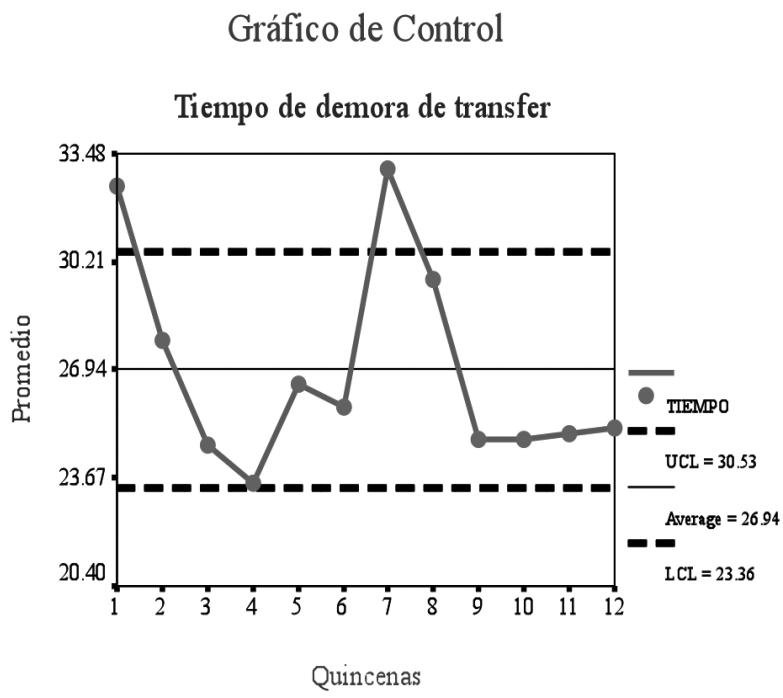
EJERCITACIÓN

Cada quince días, la Agencia de Viajes Y realiza como promedio seis transfers desde el aeropuerto provincial hasta el Hotel A, con clientes que arriban a dicha instalación. Cada uno de los transfers dura aproximadamente el mismo tiempo, pero en los últimos seis meses, se han notado algunas variaciones en la duración del viaje, y se piensa que las mismas, están provocando descontrol en el proceso de transportación, y por ende, desorden en los horarios de check-in al hotel.

La agencia va a comprobar, si ha existido alguna variabilidad notable que haya afectado la calidad del servicio de transfer de llegada. Los datos para el análisis se brindan a continuación:

	Tiempo de demora de los transfers de llegada durante seis meses											
Transfers	**15 Ene**	**30 Ene**	**15 Feb**	**28 Feb**	**15 Mar**	**30 Mar**	**15 Abr**	**30 Abr**	**15 May**	**30 May**	**15 Jun**	**30 Jun**
1	35	30	27	23	27	23	35	30	25	29	25	24
2	29	25	21	27	28	25	32	35	23	24	27	26
3	31	27	23	22	29	23	38	33	22	26	23	28
4	39	26	27	23	25	27	29	29	27	26	21	23
5	33	34	28	22	26	29	33	24	28	23	28	25
6	28	25	22	24	24	28	31	27	24	21	26	25

SOLUCIÓN



La Agencia de Viajes Y puede afirmar que el proceso de transportación no se ha comportado de manera estable, puesto que ha habido dos días fuera de control. Estos fueron el 15 de enero y el 15 de abril. Habría entonces que revisar, cuáles han sido las causas de tales variaciones en ambos días.

Bibliografía.

Guerra, Caridad et al. **"Estadística"**. Tercera edición. Editorial Félix Varela. La Habana, 2004.

Gutiérrez Pulido, Humberto y de la Vara Salazar, Román. **"Control estadístico de la calidad y seis sigma"**. Volumen 1. Editorial Félix Varela. La Habana, 2007.

Hernández Sampieri, Roberto. **"Metodología de la investigación"**. Segunda edición, 2003.

Levine, David et al. **"Statistics for managers using Microsoft Excel"**. Segunda edición. Editora Prentice-Hall. Brasil.

Prieto Valiente, Luis y Herranz Tejedor, Inmaculada. **"¿Qué significa estadísticamente significativo?"**. Editorial Díaz de Santos S.A. España, 2005.

Pupo, Juana et al. **"Análisis de regresión y series cronológicas"**. Tercera edición. Editorial Félix Varela. La Habana, 2004.

Santos Peñas, Julián y Muñoz Alamillos, Ángel. **"Fundamentos de estadística aplicados al turismo"**.