



**UNIVERSIDAD NACIONAL
AUTÓNOMA DE MÉXICO
FACULTAD DE ESTUDIOS SUPERIORES
ACATLÁN**

**PORQUE EL INSTINTO Y LOS LIBROS NO SON SUFICIENTES...
SPSS UNA HERRAMIENTA DE SOFTWARE PARA EL
APRENDIZAJE DE LA ESTADÍSTICA**

T E S I S

**QUE PARA OBTENER EL TÍTULO DE
LICENCIADA EN MATEMÁTICAS APLICADAS Y
COMPUTACIÓN**

P R E S E N T A

GONZÁLEZ MALDONADO EDITH CAROLINA

**ASESORA: DRA. MARÍA DEL CARMEN GONZÁLEZ
VIDEGARAY**

OCTUBRE, 2011



Universidad Nacional
Autónoma de México



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Agradecimientos

Mamá y Papá, por cuidarme en salud y en enfermedad, por haberme enseñado a ser fuerte y valiente, a no dejarme de nadie, a creer en Dios y en mí misma, y a valorar lo que tengo.

Ricardo, por ser un hermano que más que cuidarse a sí mismo cuidar de nuestra familia; porque aún en la debilidad encontraste fortaleza.

Bety, por ser mi hermana favorita, por esos consejos, por cuidarme y procurarme a mí y también a nuestros padres, sobre todo por ser un ejemplo a seguir.

Luis, por esos consejos y escuchar mis peroratas, además de hacerme ver mis errores y darme las posibles soluciones.

A mis amigos de la primaria, secundaria y preparatoria; Itzel, Miguel Beltrán, Marco Jardón, Lucía, Mariana, Daniela, Reyna, Erika, Nayeli; Miriam, Sonia, Rosa, Mariana; porque me enseñaron a dar todo por la amistad y por sus buenos deseos.

A los profesores Anabel Moreno, Guadalupe Moreno, Nasheli López, Christian Delgado, Beatriz Trueba, Claudia Sierra, Carlos Hernández, Ricardo Domínguez, Sandra Ortiz, Luz María Lavín, Pablo González, Jaime Vergara; por haberme demostrado que es posible dar un 1,000% a pesar de las adversidades y el miedo a fracasar.

Dra. MariCarmen González Videgaray, por apoyarme en este gran reto, por despertarme el interés hacia la estadística, así como abrirme el panorama hacia un mundo lleno de conocimientos y novedades tan fascinantes, y por iniciarme a la aventura de escribir.

Mtra. Araceli Álvarez, no tengo palabras para agradecer haber creído en mí, y haberme adentrado a la enseñanza del álgebra, por tenerme la confianza y tratarme como amiga, entenderme y celebrar conmigo mis pasos por Acatlán.

Fer, mi querido Fer, por ser mi profesor, amigo, hijo madriño; nunca creí que llegaras a ser tantas cosas para mí, doy gracias porque estuviste, estás y estarás en mi vida, compartiéndome tu sabiduría, tu amistad, tu carisma y tu cariño.

Al Área de Cursos de la FES Acatlán, Inge, Gabo, Angy, Eric chino, Lupita; por haberme ayudado en varios momentos.

Angel, porque cada día me enseñaste algo nuevo, además, de la nada nació una bella amistad de la que espero nunca desprenderme, te quiero mucho amigo.

Al Taller de Álgebra Elemental, mis niños del 1151 y 1154 de cuatro generaciones; gracias a todos, ya que me impulsaron a esmerarme y motivarlos a enamorarse de la hermosa carrera de Matemáticas Aplicadas y Computación.

Raúl, amigo, gracias por estar ahí, siempre dispuesto a escuchar mis quejas, aconsejarme y alentarme siempre a continuar con mis planes y mucho más.

Lupe, miya, por tu amistad, paciencia, confianza y apoyo en todos estos años.

Richí, ¡¡Manito!! No sabes cuántas veces fuiste de gran ayuda y ahí estuviste, sé que llegaremos lejos por la amistad.

A mis amigos de la carrera: Santos, Felipe, Adrián, Gus, Neto, Minerva, Mireya, Pancho, Roy, Tío Jou, Gabby, Yadira, Jessy; porque compartimos grandes experiencias que nos unieron como personas.

Isis, porque de la nada te convertiste en una amiga, por hacerme ver que si sigo aquí es por algo; y sobre todo por apoyarme y dejarme conocer a la esa persona humilde, luchadora, sonriente que eres.

A mis amigos de la Auditoría Superior de la Federación: Vero, Cynthia Elizabeth, Oscar Ordoñez, Oscar Miranda, Gildardo, Cynthia Nava, Nayelí, Alejandro, Lety, Ivonne, Pablo, Jesús, Mike, Mike Otho, Alfonso, Demián, Lilián; por estar ahí apoyándome en mis retos.

Finalmente, a la que fue y seguirá siendo mi segunda casa, a la que me dio el orgullo de ser Acatlense... por eso, gracias UNAM...

¡México, Pumas, Universidad!

Goya, goya, cachún, cachún ra ra

cachún cachún ra ra goya...

¡¡¡UNIVERSIDAD!!!

ÍNDICE TEMÁTICO

Introducción	- 9 -
Justificación	- 11 -
¿Para qué la estadística?	- 11 -
¿SPS... qué?	- 12 -
Capítulo 1. Mi primer acercamiento a SPSS	- 13 -
1.1 Ficha: Vista de datos	- 15 -
1.2 Ficha: Vista de variables	- 15 -
1.3 Introduciendo datos	- 17 -
1.4 Tipos de archivo SPSS	- 18 -
1.5 Algunas operaciones	- 18 -
1.5.1 Calcular variable	- 18 -
1.5.2 Recodificación	- 19 -
1.5.3 Agrupación visual	- 21 -
1.5.4 Reemplazar valores perdidos	- 22 -
1.6 Manejo de los casos	- 22 -
1.6.1 Ordenar casos	- 23 -
1.6.2 Segmentar archivo	- 23 -
1.6.3 Filtrado (seleccionar casos)	- 23 -
1.6.4 Ponderar casos	- 24 -
1.7 Fundir archivos	- 25 -
1.8 Salvando nuestro trabajo	- 25 -
1.9 Ejemplo general	- 26 -
Capítulo 2. Estadística descriptiva	- 29 -
2.1 Medidas de tendencia central	- 29 -
2.2 Medidas de dispersión	- 31 -
2.3 Medidas de forma	- 31 -
2.4 Medidas de tendencias no central o de posición	- 33 -
2.5 Gráficos	- 33 -

2.5.1	Gráfico de dispersión simple	- 33 -
2.5.2	Histograma	- 33 -
2.5.3	Gráfico de barras	- 33 -
2.5.4	Gráfico de caja	- 34 -
2.6	Ejemplo general	- 34 -
2.7	Ejercicio complementario	- 41 -
2.8	Caso de estudio	- 45 -
Capítulo 3.	Pruebas de hipótesis	- 47 -
3.1	Comparación de medias	- 48 -
3.1.1	Medias	- 48 -
3.1.2	Prueba T para una muestra	- 49 -
3.1.3	Prueba T para muestras independientes	- 51 -
3.1.4	Prueba T para muestras relacionadas	- 54 -
3.1.5	ANOVA de un factor	- 55 -
3.2	Pruebas no paramétricas	- 58 -
3.2.1	Prueba de chi-cuadrado	- 58 -
3.2.2	Prueba binomial	- 60 -
3.2.3	Prueba de rachas	- 60 -
3.2.4	Prueba de K-S para 1 muestra	- 61 -
3.2.5	Pruebas para 2 muestras independientes	- 62 -
3.2.6	Pruebas para K muestras independientes	- 65 -
3.2.7	Pruebas para 2 muestras relacionadas	- 66 -
3.2.8	Pruebas para K muestras relacionadas	- 68 -
3.3	Ejemplo general	- 70 -
3.4	Ejercicio complementario	- 75 -
3.5	Caso de estudio (Continuación...)	- 76 -
Capítulo 4.	Correlación y regresión lineal	- 79 -
4.1	Correlación	- 79 -

4.1.1	Correlaciòn bivariada	- 79 -
4.1.2	Correlaciòn parcial	- 81 -
4.2	Regresiòn lineal	- 81 -
4.2.1	Regresiòn lineal simple	- 82 -
4.2.2	Regresiòn lineal múltiple	- 85 -
4.3	Ejemplo general	- 85 -
4.4	Ejercicio complementario	- 88 -
4.5	¿Otro ejercicio complementario? Por qué no...	- 89 -
4.6	Caso de estudio (Continuaciòn...)	- 91 -
Capítulo 5.	Series de tiempo	- 92 -
5.1	Funciòn de autocorrelaciòn y de autocorrelaciòn parcial	- 94 -
5.2	Periodograma	- 95 -
5.3	Modelos estacionarios y no estacionarios	- 95 -
5.3.1	Modelos autoregresivos y de medias móviles	- 96 -
5.3.2	Modelos ARIMA	- 98 -
5.4	Suavizamiento exponencial	- 100 -
5.4.1	Modelos no estacionales	- 100 -
5.4.2	Modelos estacionales	- 101 -
5.5	Descomposiciòn estacional	- 101 -
5.6	Crear serie temporal	- 102 -
5.7	Ejemplo general	- 103 -
5.8	Ejercicio complementario	- 105 -
	Conclusiones	- 107 -
	Fuentes de informaciòn	- 109 -
	Anexo A. Sintaxis de SPSS	- 111 -
	Anexo B. ¿Qué hay de nuevo?... SPSS 19	- 113 -
	Índice de Figuras	- 114 -
	Índice de Tablas	- 116 -

INTRODUCCIÓN

Hoy en día la interacción de varias disciplinas se ha presentado en más de un país, aún más interesante son las disciplinas que se apoyan de otra u otras más para dar mayor alcance a su estudio, a su aplicación, etc.; esto a su vez está rompiendo con el tan famoso mito que la gente ha tenido, esto es, que se va estudiar algo completamente opuesto por alejarse de una disciplina. El caso más común es cuando la gente por temor a las matemáticas estudia leyes, política, química, biología; pero una vez que se incorpora al campo laboral puede llegar a necesitar alguna herramienta matemática.

Este mito no es exclusivo de unas cuantas universidades o de un país, y pudo haberse debido por la forma en que fueron enseñadas las matemáticas, es decir, basándose exclusivamente en la teoría y ejemplos poco ilustrativos; es por ello la hipótesis de que si la estadística para los alumnos resulta muy teórica y tediosa debido a la realización de análisis como obtener promedios, encontrar valores en tablas, regresiones lineales; entonces el uso de un software estadístico para aplicaciones prácticas reales facilitará el aprendizaje e interpretación de la estadística.

El objetivo de esta tesis es: Facilitar el aprendizaje e interpretación de la estadística para motivar el estudio de la misma, con el apoyo del software estadístico SPSS.

Para cumplir con el anterior objetivo, esta tesis se ha dividido en: Justificación, cinco capítulos, conclusiones, fuentes de información, y anexos.

La dinámica de la presente tesis, consiste en que el lector no sólo vea teoría, sino más bien en ejemplificar lo visto en cada capítulo de tal forma que no nos quedemos con la sensación de que sólo son matemáticas, ya que cada uno de los ejemplos fue tomado de la realidad; ya sea con datos del Instituto Nacional de Estadística y Geografía (INEGI), o del Banco de México, o de alguna otra página de la que podamos obtener datos confiables y manejables de acuerdo a lo visto en el capítulo respectivo.

Además, para probar la efectividad de este material, se desarrolló un pequeño experimento, el cual es presentado a lo largo de la tesis.

Pues bien, empecemos a conocer este software y aprendiendo, no, mejor dicho a entender y saber interpretar los resultados obtenidos.

JUSTIFICACIÓN

¿Cuántas veces nos hemos preguntado por qué tenemos que hacer algo?, o también nos hemos planteado ¿de qué sirve aprender tal cosa?; pues bien en este apartado veremos los fundamentos del por qué “debemos” aprender estadística, y para ser más precisos, aprender a interpretarla.

Podremos decir que si bien la estadística tiene que ver con matemáticas, esta no es exclusiva de matemáticos, ya que debido a su alcance, y de acuerdo con Behar [6], en varias licenciaturas se han incluido al menos un curso de estadística dentro de su plan de estudios, para esto bastará mencionar que licenciaturas como Ciencias Políticas y Administración Pública, Economía, Relaciones Internacionales, Sociología, Comunicación, Ingeniería Civil, Actuaría y, Matemáticas Aplicadas y Computación, todas ellas impartidas en la Facultad de Estudios Acatlán.

Además, dentro de los programas de asignatura de las materias relacionadas con la estadística se sugiere el uso de software como apoyo didáctico, así como fomentar en el alumno la investigación relacionada con la materia y con temas relevantes que se encuentren en revistas especializadas.

Por lo anterior se debe involucrar al estudiante con temas reales que puedan ser analizados e interpretados mediante la estadística, de este modo se puede fomentar tanto la interpretación así como la investigación.

Asimismo, a lo largo de esta tesis identificaremos que el manejo de un software estadístico facilita el quehacer del análisis estadístico, ya que debido a la automatización de estadísticos permite un mayor enfoque al análisis.

¿PARA QUÉ LA ESTADÍSTICA?

Como ya habíamos mencionado, la estadística no es exclusiva de matemáticos hoy en día a nivel mundial se ha convertido en una de las herramientas más socorridas para la toma de decisiones así como para la elaboración de proyectos de investigación [17].

La problemática del aprendizaje de la estadística no es novedad, ni tampoco un fenómeno que sea propio de un solo país. De acuerdo con Behar [5] se debe principalmente al enfoque con el que se transmiten los conceptos, ya que el método pedagógico ha tenido un enfoque cuantitativo, es decir, un buen estudiante es aquel que aprende mucho, sin embargo, no se hace hincapié en la integración de la teoría con la conceptualización de la vida real.

Lo anterior sería en el mejor de los casos, en los que el estudiante muestra interés y aprende lo que el profesor le enseñe, sin embargo, la mayoría se encuentra frustrado debido a que el contexto es basado en números y en la obtención de ellos mediante diversos cálculos, y no en la interpretación de los resultados, de este modo a lo que se desea llegar es a un enfoque cualitativo [5] en el que no sólo se maneje teoría, sino que el estudiante construya su propio conocimiento mediante la conexión de éste con otros temas que sean de su interés.

Es importante resaltar la importancia que ha tomado el uso de software y tecnologías de la información para la enseñanza, de este modo, las ventajas de usar software son: la automatización de procesos, la optimización del tiempo para realizar cálculos, en general facilita al usuario alguna tarea.

A todo esto, Behar [5] propone, lo mismo que Biggs, esto es:

- Un contexto motivacional positivo. Permitirá al estudiante resolver problemas que tenga, para generar motivos auténticos para aprender estadística.

- Grado de actividad por parte del estudiante. Las mejores preguntas son del estudiante, de este modo el estudiante resuelve sus dudas, ya que después de todo el profesor no tiene todo el conocimiento ni la verdad absoluta.
- Interacción con otros estudiantes. Se refiere a la confrontación del conocimiento con otra opinión y llevar así a una conclusión.
- Base de conocimientos adecuada. Está involucrada una adecuada estructura del conocimiento y la relación con el contexto necesario.

¿Qué tan bueno resulta separar a la estadística de las matemáticas? Indudablemente las bases matemáticas permiten establecer los fundamentos teóricos y metodológicos de la estadística; sin embargo, limita a los estudiantes que no les gustan, por lo que sería más factible que más que separarlas se haga una mezcla de deducción con inducción, en la que se involucre tanto la teoría así como el análisis de ejemplos reales.

¿SPSS... QUÉ?

SPSS (*Statistical Product and Service Solutions*) es un software estadístico comercial que contempla una gran variedad de análisis predictivo, es eficiente y fácil de usar para la organización y análisis de datos; además, es muy usado dentro del mundo de los negocios y de la educación en más de 100 países [11].

La cobertura de clientes de SPSS abarca las áreas como: educación, servicios financieros, gobierno, cuidado de la salud, estudios de mercado, telecomunicaciones; dentro de las cuales los análisis más utilizados son: minería de datos, modelos de series de tiempo, y control de calidad.

Cabe resaltar que SPSS es el nombre que tenía antes de la versión 17; ya que a partir de esta fue renombrado como PASW Statistics.

Las principales características que destacan de las versiones 13 a 18 son: el manejo de la interfaz, el idioma y sistema operativo en que puede ser instalado; ya que la versión 18 la podemos instalar en sistemas operativos: MAC, Linux y Windows. Además, admite el uso de procedimientos escritos en R¹ y scripts de Python², permite personalizar cuadros de diálogo; así como una mayor gama de análisis y gráficos.

Finalmente, podremos preguntarnos, ¿por qué no usar alguna alternativa de software libre? Pues bien, el software libre para estadística nos permite realizar cálculos de los diferentes estadísticos, sin embargo, este no contiene la misma gama de herramientas estadísticas que ofrece el software comercial tal como SAS, SPSS, Arena, Statgraphics, entre otros.

El software libre más conocido en el ámbito de la estadística es R, el cual es un lenguaje de programación con el que se pueden obtener técnicas estadísticas y gráficos, incluyendo modelos lineales y no lineales. Cabe destacar que R no maneja cuadros de diálogo, por lo que el usuario debe de conocer primero la sintaxis, estructura e instrucciones que maneja el lenguaje de programación, ya que no podríamos manipular todas las herramientas que posee.

¹ Software estadístico libre.

² Lenguaje de programación orientado a objetos que permite trabajar más rápido e integrar sistemas.

Capítulo 1. MI PRIMER ACERCAMIENTO A SPSS

El ordenador ha sido hasta ahora el producto más genial de la vagancia humana.

Slogan de IBM

OBJETIVO: En este capítulo manejaremos las funciones principales de SPSS, para que conozcamos el entorno del programa.

La versión que usaremos es la de nombre “PASW Statistics 18”, la cual es una versión anterior a la más reciente (a lo largo del documento la nombraremos como SPSS).

Una vez instalado el programa, al ingresar por primera vez saldrá el cuadro de diálogo de la Fig. 1, éste nos permitirá elegir entre ejecutar el tutorial, ingresar manualmente los datos, realizar una consulta, crear una nueva consulta, abrir un archivo existente (ya sea de SPSS o de otro origen). Dicha ventana se abrirá cada vez que ingresemos al software, de modo que al marcar en la casilla “No volver a mostrar este cuadro de diálogo”, se dejará de mostrar en ocasiones posteriores.

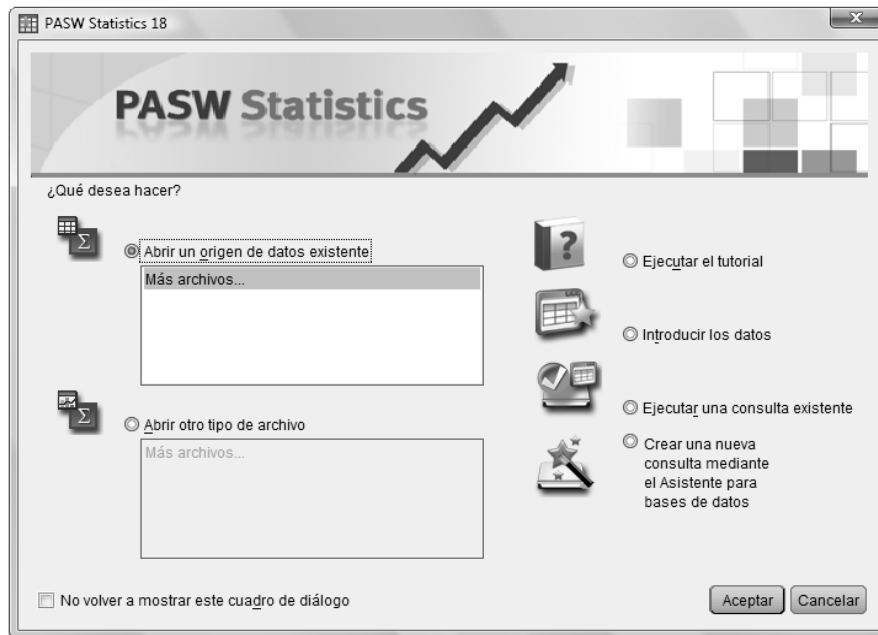


Fig. 1. Asistente para introducir datos

SPSS maneja simultáneamente dos tipos de ventanas, las cuales son:

- Editor de datos (ver Fig. 2): En esta ventana podemos visualizar los datos.
- Visor de resultados (ver Fig. 3): En esta ventana SPSS coloca los resultados de los cálculos elegidos.

En ambas ventanas podemos realizar las mismas acciones, excepto que los menús de Insertar y Formato no los encontraremos en la ventana del editor de datos, ya que son opciones propias de la ventana del visor de resultados.

SPSS predetermina un idioma para todo su entorno; sin embargo, podemos cambiarlo siguiendo la ruta: Edición → Opciones → General, en la sección de “Resultado” y en la de “Interfaz” seleccionamos el idioma preferido.

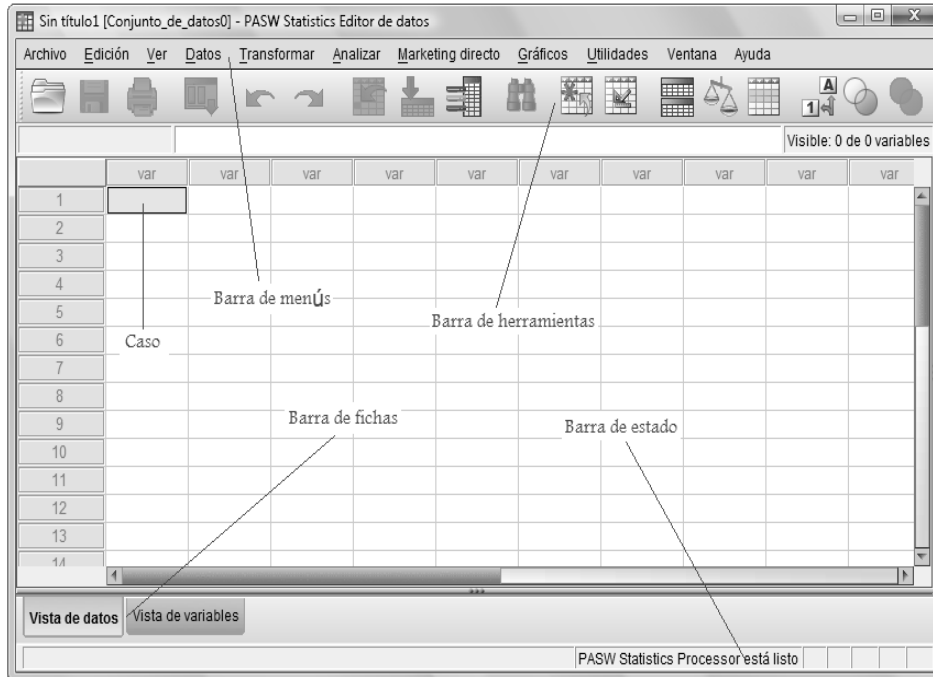


Fig. 2. Editor de datos

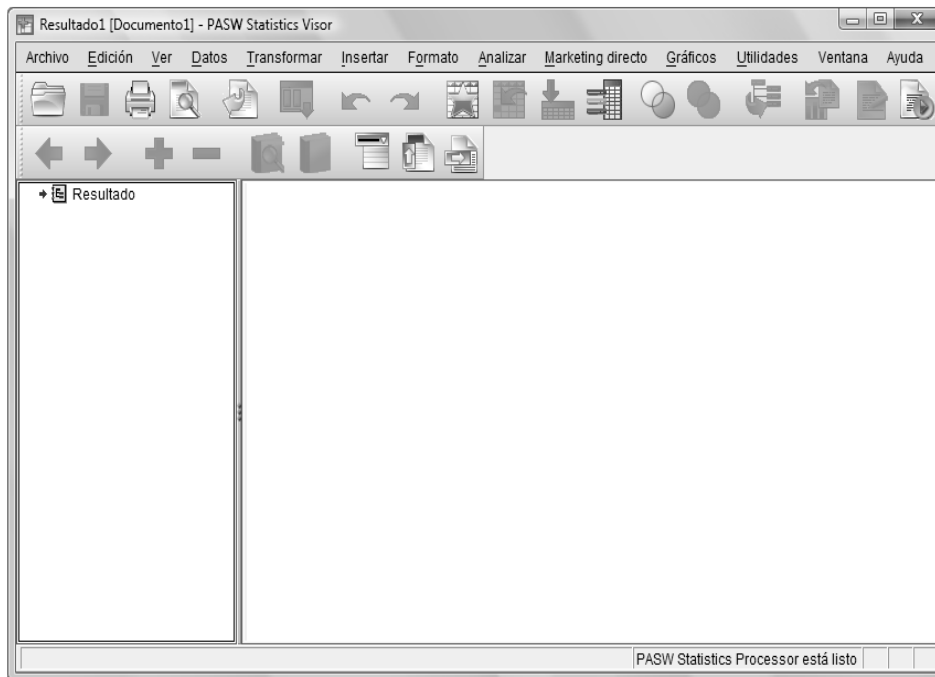


Fig. 3. Visor de resultados

1.1 FICHA: VISTA DE DATOS

En esta ficha podemos ver los datos tal cual los ingresamos, no importando si son numéricos, alfabéticos o alfa-numéricos.

Cada columna corresponde a una variable, la cual será nombrada por default como “VAR0001” al momento en que ingresemos algún dato. Si colocáramos otro dato en la siguiente columna el nombre que recibiría es el de “VAR0002”, y así sucesivamente.

El número de renglones irá aumentando conforme tengamos más datos. A la intersección de renglón y columna le llamaremos CASO, y estos serán por cada variable.

1.2 FICHA: VISTA DE VARIABLES

La vista de variables nos permitirá establecer los atributos de cada una de ellas; tales como el nombre, la longitud, el tipo de variable, alineación. Los atributos dependerán del tipo de variable, estas pueden ser:

1. Cualitativas, tienen que ver con cualidades, es decir, aspectos que no se pueden medir numéricamente. Por ejemplo, variables que midan el género, el nivel socioeconómico, alguna categoría de trabajo.
2. Cuantitativas, es posible ponerlas en correspondencia con algún número. Por ejemplo, variables que midan la edad, el peso, la estatura, Coeficiente Intelectual.

Con la columna “**Nombre**”, podremos nombrar a la variable tomando en cuenta las consideraciones descritas a continuación:

- Cada nombre debe ser único.
- La longitud puede ser de hasta 64 bytes, sin usar espacios.
- El primer carácter debe ser una letra o alguno de estos caracteres: @, # (para variables transitorias), o \$ (variable del sistema). El punto no está permitido.
- No usar “_” al final del nombre.
- Las palabras reservadas no están permitidas, tales como: ALL, AND, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO y WITH.
- Hay diferencia entre mayúsculas y minúsculas.

Usando el atributo “**Tipo**”, definiremos el tipo de variable de la que se trate. SPSS por default define cada variable nueva como numérica, para cada tipo nos pedirá el formato que necesitemos.

Los tipos de variables que contempla SPSS son:

1. Numérico: Se refiere a datos tipo número, ya sea en formato estándar o científico.
2. Coma: Es una variable numérica que será delimitada cada tres dígitos por una coma.
3. Punto: Muy parecido al anterior pero el separador será un punto y la coma separará los dígitos decimales.
4. Notación científica: Mostrará una E intercalada y un exponente con signo como potencia de base 10.


5. Fecha: Es el indicado cuando tenemos variables que miden el tiempo. La fecha puede ser con respecto a meses, años, días, trimestres, e incluso manejar formatos de horas.
6. Dólar: Añade el signo de dólar al inicio, además de usar la configuración del tipo “punto”.
7. Moneda personalizada: Se trata de una variable numérica con algún formato de moneda que personalizemos usando la anchura y los decimales.
8. Cadena o alfanumérica: Este tipo de variable considera cualquier letra o combinación con números.

Las configuraciones hechas en cuanto a presentación no influyen en la forma en que ingresamos los datos, es decir, si introducimos un 1 y se le da formato de Dólar SPSS automáticamente pondrá el símbolo correspondiente.

La “**Anchura**” se refiere al número total de dígitos, incluyendo las cifras decimales que tendrá el dato en caso de ser numérico, tiene como límite 16 dígitos.

Si deseamos ponerle un nombre muy descriptivo a la variable, entonces debemos emplear la opción “**Etiqueta**”, la cual admite hasta 256 caracteres, permite usar espacios; pero, ¿qué nombre aparecerá para la variable?, en la vista de datos aparecerá el que le dimos en “Nombre”; sin embargo, cuando se genere el resultado de alguna operación, serán reportados con el nombre que le dimos en la “Etiqueta”.

Existe otro tipo de etiqueta, esta es la “**Etiqueta de valor**”, la cual nos servirá para darle cualidades a las variables que no tienen sentido por sí solas. Por ejemplo, en la variable “Género” se capturaron los valores unos y ceros, en los que el uno se referirá a masculino y el cero a femenino; esto lo especificaremos mediante las Etiquetas de valor.

Observemos la Fig. 4, en la casilla de valor pondremos el valor que se vaya a describir, para la casilla de etiqueta debemos introducir la cualidad, de esta forma aparecerá el valor pero no la descripción, si queremos que aparezca basta con que presionemos el botón , que es el de “Etiquetas de valor”, y ¡listo!

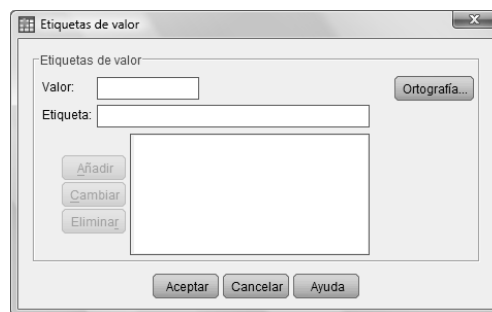


Fig. 4. Etiqueta de valor

Habrán veces en las que no se tendrán todos los datos completos o correctos, ya sea porque es un valor nulo o porque fueron mal capturados, para estos casos tenemos el atributo de “**Perdidos**”, el cual nos permite definir “valores perdidos” (aquellos que no se tomarán en cuenta para los cálculos) mediante dos formas:




1. Eligiendo de uno a tres valores que no sean correctos para la información que estemos capturando.
2. Usando un intervalo de valores, esta opción también nos permitirá dar un valor además de usar el rango de valores que SPSS tomará como perdidos.

No se consideran valores perdidos a los valores de cadena, incluso vacíos o nulos; a no ser que se configure de esa forma.

“**Columnas**”, nos servirá para establecer la anchura de la columna, aunque también podemos usar los bordes de la vista de datos. Los cambios no implican modificar el ancho definido para la variable.







“**Alineación**” nos permitirá dar el formato de alineación a la izquierda, derecha y centrado.

Con la columna “**Medida**” clasificaremos a la variable, para lo cual contamos con las medidas siguientes:

-  Escala: Mide alguna observación, tal como la edad en años, ingresos percibidos.
-  Ordinal: Es posible ordenar las variables, y estas representan categorías, por ejemplo, niveles de satisfacción.
-  Nominal: Las variables no poseen un orden natural, pero si representan categorías, por ejemplo: regiones, código postal.

Todos los atributos descritos anteriormente también podemos configurarlos usando: Datos → “Definir propiedades de las variables”, en este cuadro de diálogo vamos analizando los atributos de cada variable.

La última versión de SPSS, anexó el atributo “**Rol**”, la cual tiene el objetivo de que podamos pre-seleccionar a las variables según sea el análisis, para ellos veamos ahora cómo define sus tipos de roles:

1.  Entrada: La variable será usada como una entrada (variable independiente), es la que SPSS predetermina.
2.  Objetivo: Recomendada para usarla como una variable de salida (variable dependiente).
3.  Ambos: Esta variable puede ser usada tanto como de salida como de entrada.
4.  Ninguna: No se le asigna un rol a la variable.
5.  Partición: La variable será usada para partir el archivo
6.  Segmentar: Ésta función la veremos con mayor detalle en el tema 1.6.2.

Con esto nuestras variables quedan descritas, pero si fuera necesario algún otro atributo, SPSS nos da la opción de crear nuestros propios atributos, para esto debemos estar en la vista de variables, dar clic en Datos → “Nuevo atributo personalizado”.

1.3 INTRODUCIENDO DATOS

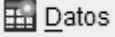
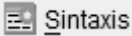
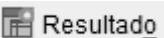
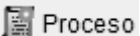
La forma más sencilla para transportar los datos a un archivo de SPSS, es usando el famoso “copy-paste”; esto lo podemos hacer con uno o más datos.

Otro camino para traernos todos los datos de un archivo es usando: Archivo → “Leer datos de archivo” (ver Fig. 5). Una vez ahí, seleccionaremos el archivo a abrir y los datos se pasarán a la hoja de trabajo, en donde para cada columna se creará una variable.

Con esto, tenemos la oportunidad de traer datos de otro tipo de extensión, distinto al del programa, siempre y cuando estén capturados como una tabla.

1.4 TIPOS DE ARCHIVO SPSS

Al igual que otros programas, SPSS maneja distintos tipos de extensiones de archivo, cada uno tiene distintas finalidades, mismas que veremos en seguida:

1.  **Datos**. La extensión **.sav** es para que guardemos los datos tal cual están capturados en la hoja de trabajo.
2.  **Sintaxis**. Como una herramienta PLUS, SPSS usa también comandos con los cuales podemos realizar las mismas acciones que en el entorno de botones; de modo que la extensión para guardar archivos de sintaxis es la **.sps**. (Para mayor información, ver Anexo A)
3.  **Resultado**. Como ya habíamos mencionado, el visor de resultados maneja su propia ventana, de forma que para guardar los resultados obtenidos usaremos la extensión **.spv**.
4.  **Proceso**. Los archivos del tipo proceso usan un lenguaje de programación para automatizar tareas, con extensiones **.wwd** o **.sbs**.

Ahora ya sabemos cuándo usar uno u otro tipo de extensión de archivo.

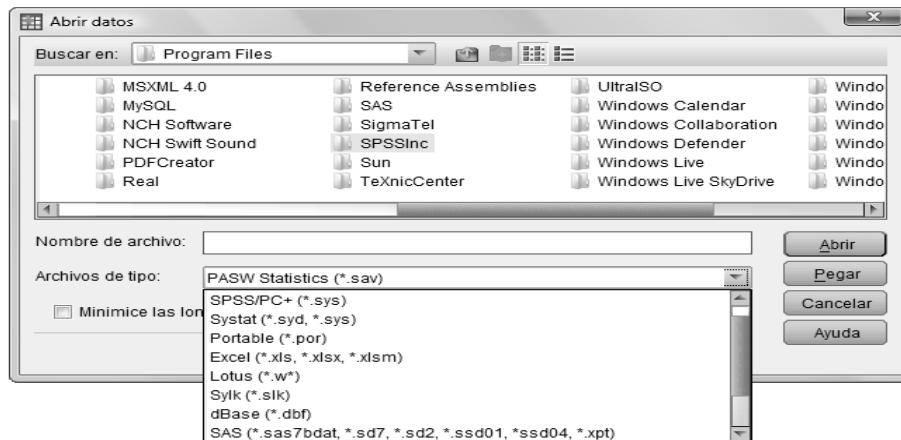


Fig. 5. Abrir datos

1.5 ALGUNAS OPERACIONES

A continuación veremos la forma en que podemos realizar operaciones con las variables. Dentro del menú “Transformar” encontraremos las opciones para hacer los cálculos con las variables además de revisar algunas de las funciones con las que cuenta SPSS.

1.5.1 CALCULAR VARIABLE

Dentro de “Calcular variable” (ver Fig. 6), podemos hacer cálculos con las variables, por ejemplo:

1. Manejo de funciones aritméticas, estadísticas, de valores perdidos, de distribuciones, de puntuación, de variables aleatorias, de cadena; y de fecha y hora. SPSS nos da una breve explicación de lo que hace cada tipo de función.

2. Condicionar el cálculo, de esta forma el cálculo se hará sólo si la variable cumple con la condición especificada.

Para escoger a la variable sobre la cual trabajaremos, o incluso la función a usar, debemos presionar el botón de selección de variables o dar doble clic a la selección (ver Fig. 6). No olvidemos dar el nombre a la nueva variable que se calculará, adicionalmente podremos darle el atributo de Tipo y de etiqueta.

La forma de entrada tanto de los números como de los operadores la podremos hacer con los botones o el teclado. Por otra parte, si deseamos que el cálculo no sea para todas las variables, con el botón “Si la opción...” podemos establecer alguna condición que deba cumplir la variable para realizar el cálculo.

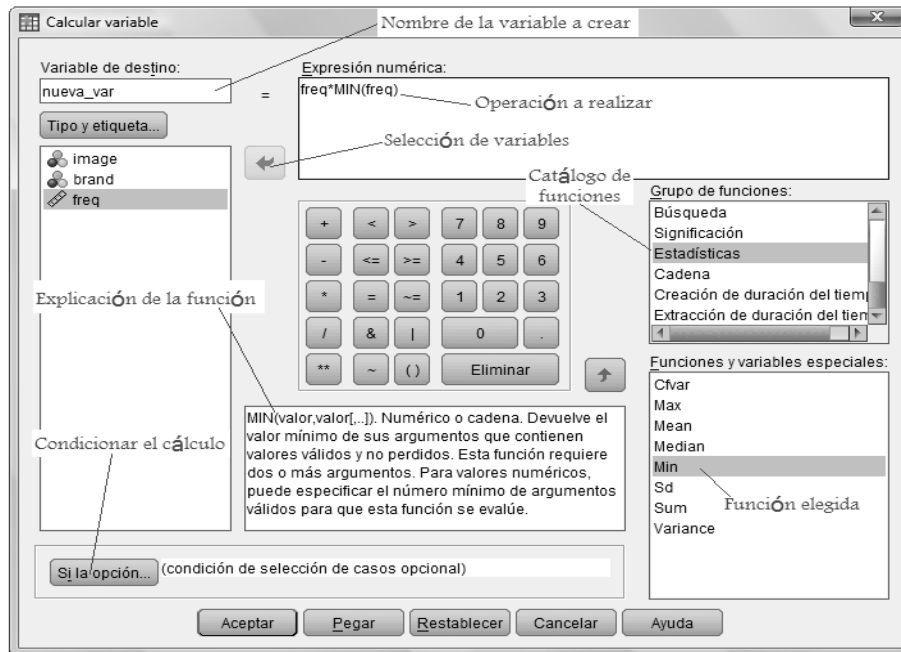


Fig. 6. Calcular variable

1.5.2 RECODIFICACIÓN

¿Cuántas veces no necesitamos replantear los valores con los cuales trabajamos? Pues bien, para ello tenemos la recodificación, la cual es una herramienta muy útil para transformar valores cualitativos a cuantitativos, esta recodificación puede ser:

1. En la misma variable, la cual al aplicarse se pierden los valores anteriores y no es posible restablecerlos. Por lo que debemos estar seguros de hacer este tipo de recodificación. El cuadro de diálogo es muy parecido al de la Fig. 7, sólo que no cuenta con la sección de los datos para la nueva variable.
2. En distinta variable, con los valores recodificados se crea una nueva variable. La Fig. 7 ilustra el cuadro de diálogo para la recodificación en otra variable.
3. Automática.

Existen diferentes criterios para recodificar, todas ellas las encontraremos al presionar el botón “Valores antiguos y nuevos”, tal como se muestra en la Fig. 8, veamos cada una de ellas:

1. Valor: Ponemos el valor que queremos cambiar.

2. Perdidos por el sistema: Cuando no hay algún valor para el dato correspondiente el programa coloca un punto en su lugar.
3. Perdido por el sistema o usuario: Se refiere a los valores que ya fueron configurados dentro de los atributos de la variable.
4. Rango: Cambia los valores que se encuentre dentro del rango que se establezca.
5. Rango, INFERIOR: Los cambios son para valores inferiores al valor que se ponga.
6. Rango, SUPERIOR: Muy parecido al anterior, sólo que ahora el límite será superior.
7. Todos los demás valores: Considera los valores que no se tomaron en cuenta con el resto de las categorías.

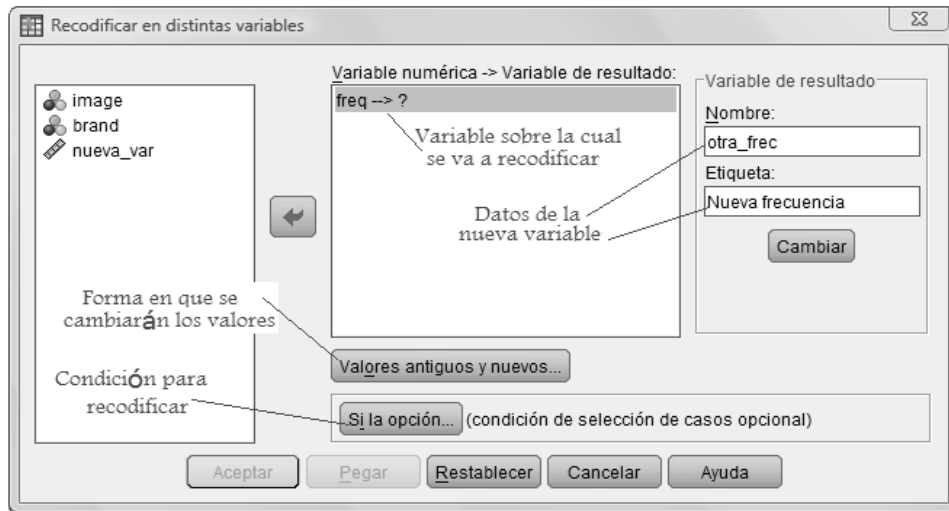


Fig. 7. Recodificar en distintas variables

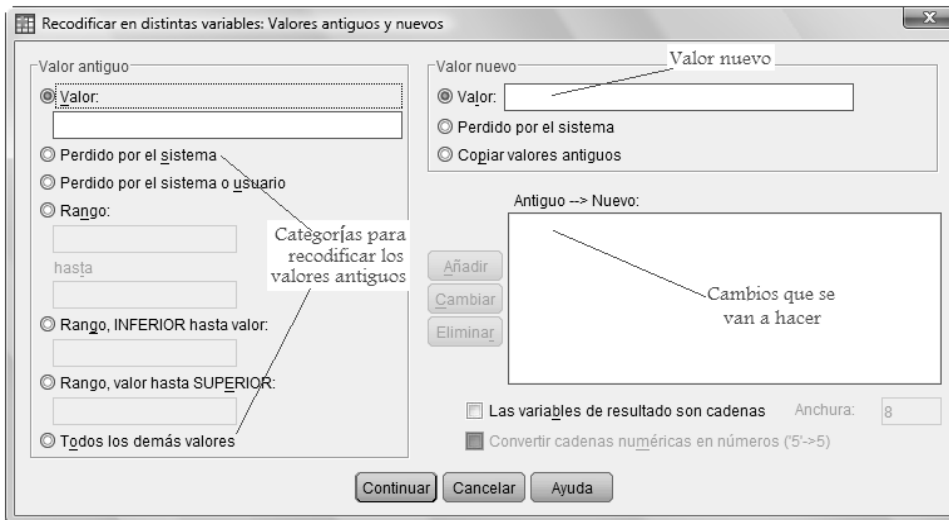


Fig. 8. Valores antiguos a nuevos

Después de que hayamos seleccionado el tipo de recodificación debemos poner el valor por el cual se va a cambiar el valor antiguo, y por último presionar el botón de añadir, de manera que podemos configurar tantas recodificaciones como nos sea conveniente.

Para el caso de la recodificación en otra variable se crea una nueva variable en la cual depositaremos los nuevos valores, para ello debemos hacer los pasos listados a continuación:

1. Elegir la variable con la que se va a trabajar, y depositarla en el área de variables. Al hacer esto aparecerá el nombre de la variable elegida seguida por una flecha y un signo de interrogación.
2. Dar un nombre a la nueva variable, y colocar una etiqueta si es necesario.
3. Presionar el botón “Cambiar”. En seguida el signo de interrogación desaparecerá y estará el nombre que hemos designado a la nueva variable.

Hecho esto debemos elegir la categoría de recodificación y dar clic en aceptar; en seguida SPSS agregará una columna para la nueva variable.

Como otra forma de recodificación que tenemos es la automática, la cual maneja el cuadro de diálogo de la Fig. 9. El proceso es parecido al de la recodificación en distintas variables, excepto que ahora debemos dar clic en el botón “Añadir nombre”, las categorías de cálculos pueden ser por “Mayor valor” o “Menor valor”.

Ambas realizan algo similar, veamos ahora la explicación:

1. Primero los valores son ordenados de menor a mayor (para el caso de “Mayor valor”) o de mayor a menor (para el caso de “Menor valor”).
2. Se hace una relación de estos valores con una secuencia de números consecutivos naturales (1, 2, 3, 4, etc.).
3. A cada nuevo valor se le coloca una etiqueta de valor correspondiente a su valor antiguo.

Luego de esto, se creará una nueva variable, y dentro de la ventana de resultados se encuentran los pasos descritos anteriormente.

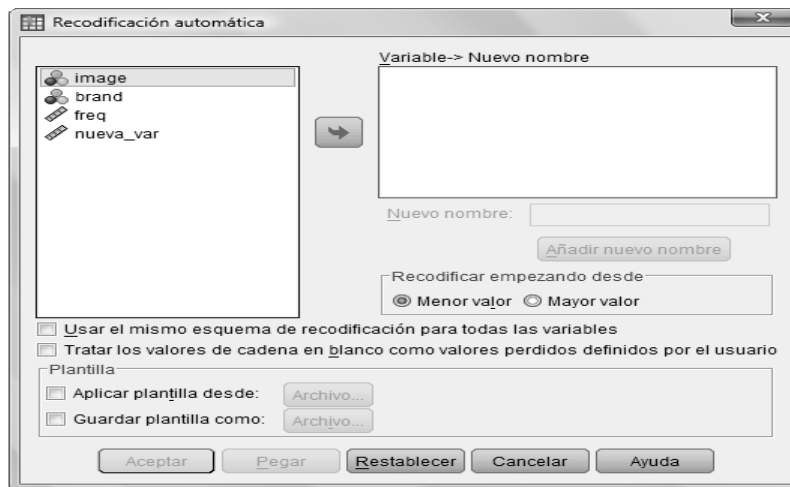


Fig. 9. Recodificación automática

1.5.3 AGRUPACIÓN VISUAL

La agrupación visual concentrará los datos de forma gráfica, su objetivo es crear variables categóricas a partir de variables de escala, esto es usando las especificaciones de la agrupación. La primera parte consiste en seleccionar las variables a agrupar, además de decidir si serán todos los casos o sólo unos cuantos. Luego de esto nos aparecerá la ventana de la Fig. 10, con ella escogeremos la variable a agrupar, sus atributos, los límites para la agrupación y las etiquetas de valor.

Cuando ingresemos un nuevo límite, la leyenda “SUPERIOR” cambiará al renglón siguiente. SPSS tiene valores establecidos para hacer las agrupaciones, esto lo encontraremos al presionar el botón “Crear puntos de corte”, que maneja las opciones:

- Intervalos de igual amplitud.
- Percentiles.
- Basados en la media y desviaciones atípicas.

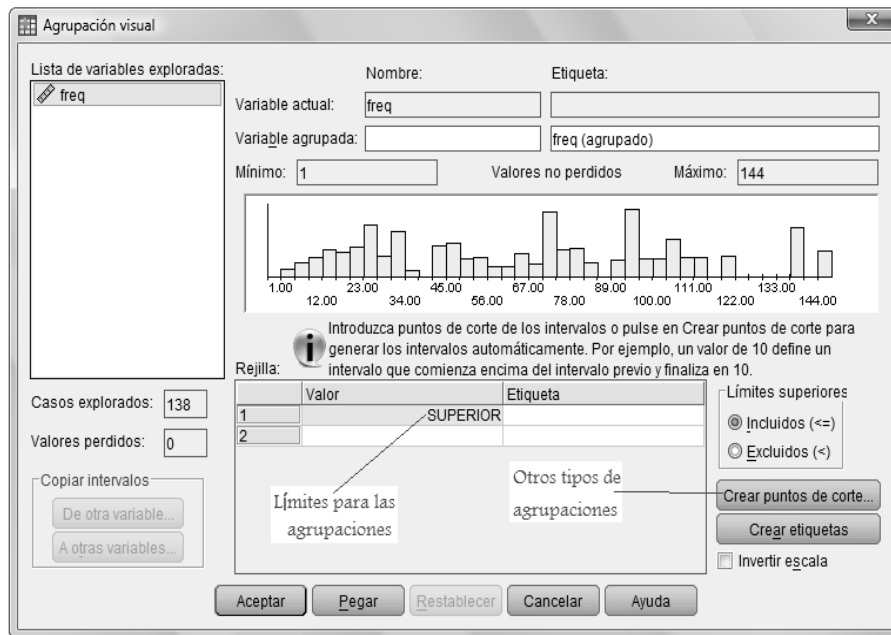


Fig. 10. Agrupación visual

Los nuevos valores que tomará la variable serán 1, 2, 3, etc.; para cada intervalo respectivamente. Con el botón “Crear etiquetas”, se harán las etiquetas en automático. Por último daremos clic en Aceptar y tendremos una nueva variable.

1.5.4 REEMPLAZAR VALORES PERDIDOS

Si dentro de nuestros datos existen valores perdidos que necesitamos reemplazar para realizar algún cálculo, esta herramienta nos permite hacerlo usando el método que prefiramos.

Una vez seleccionada la variable, el método por default es el de la media; sin embargo, contamos con otros, los cuales son: media de puntos adyacentes, mediana de puntos adyacentes, interpolación lineal, y tendencia lineal en el punto.

De tal forma que si en lugar de usar la media queremos el método de interpolación lineal, lo elegimos y presionamos el botón “Cambiar”, por último damos clic en Aceptar, enseguida se creará otra variable con los valores reemplazados.

1.6 MANEJO DE LOS CASOS


Las funciones anteriores están relacionadas con al manejo de los datos, y las funciones correspondientes a los casos son: ordenar, segmentar, filtrar o ponderar, todas ellas las podremos hacer dentro del menú “Datos”, o usando los botones de la barra de herramientas.

1.6.1 ORDENAR CASOS

Al exportar nuestros datos el programa respetará el orden con el que se encuentran, sin embargo, quizá sea necesario tenerlos ordenados.

Como ya habíamos dicho, esto lo podremos hacer usando el menú Datos → "Ordenar casos", pero también contamos con el menú contextual, ya que al seleccionar la variable y dar clic derecho aparecerá la opción de "Ordenar ascendentemente" y "Ordenar descendentemente".

1.6.2 SEGMENTAR ARCHIVO

El resultado de segmentar un archivo (botón ) es la organización de los datos con base en una variable, por lo general cualitativa, a partir de la cual se verán afectadas el resto de las variables. En la Fig. 11 tenemos el cuadro de diálogo para hacer la segmentación, para cada bloque sólo podemos elegir una opción.

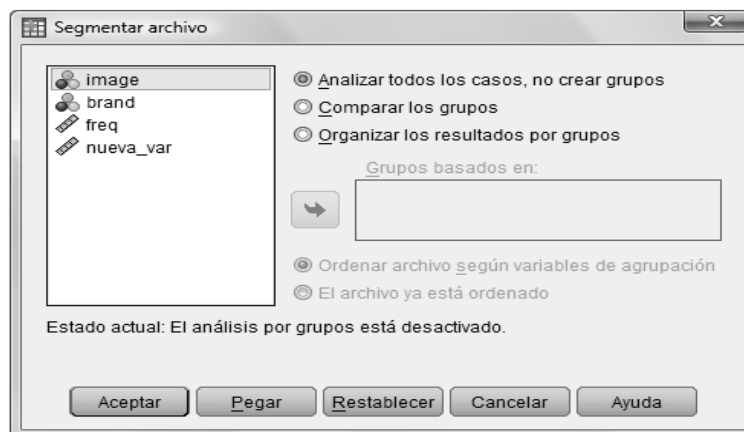


Fig. 11. Segmentar archivo

Primer bloque:


1. Analizar todos los casos. No hace ninguna segmentación.
2. Comparar los grupos. Ordena los datos con base en una variable de agrupación.
3. Organizar los resultados por grupo. Además de hacer la organización, cuando obtengamos algún resultado en el Visor, estos aparecerán organizados con base en la variable de agrupación.

Segundo bloque:

1. Ordenar archivo según variables de agrupación. Cambia el orden de los datos con base en la segmentación que se hizo.
2. El archivo ya está ordenado. Respetar el orden en el que se encuentra el archivo.

Una vez que hayamos configurado esta parte daremos clic en Aceptar y el archivo sufrirá los cambios que hemos hecho.

1.6.3 FILTRADO (SELECCIONAR CASOS)

Esta herramienta también es llamada "seleccionar casos" (botón ) , la cual nos permitirá excluir casos según nos convenga, para ello contamos con las categorías (Ver Fig. 12) enlistadas a continuación:

1. Todos los casos: Incluye a todos los casos, nos sirve para quitar el filtro si es que ya tenemos alguno.
2. Satisface la condición: Usa operadores lógicos para verificar si el valor cumple con la condición.
3. Muestra aleatoria de casos: Hace una muestra basada en porcentaje o X casos de los primeros Y posibles.
4. Basado en el rango del tiempo o de los casos: Selecciona del caso X al Y.
5. Usar variable de filtro: La variable debe de ser binaria, y así filtra los ceros o celdas vacías.

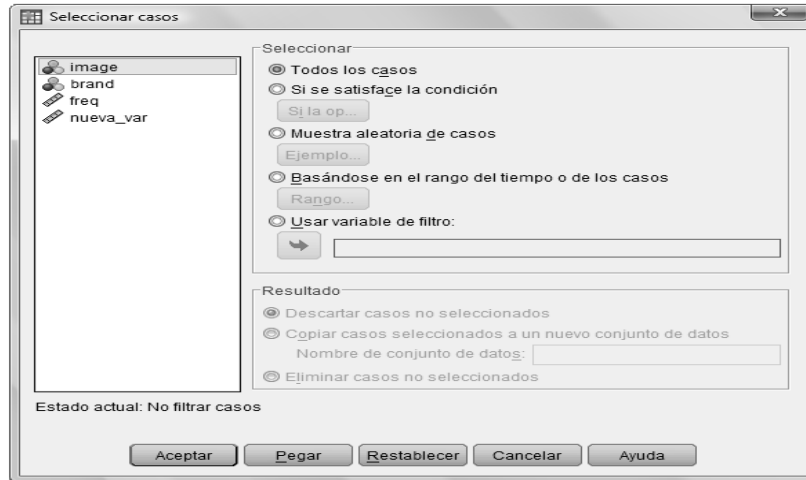



Fig. 12. Seleccionar casos

Para el resultado del filtrado podemos:

- Descartar los casos no seleccionados: Pone una diagonal en el número del caso.
- Copiar casos seleccionados a un nuevo conjunto de datos: Crea un nuevo archivo con los datos filtrados.
- Eliminar casos no seleccionados: Quita los datos que son filtrados.

Después de configurar estas opciones, se agregará una nueva variable llamada “filter_\$” que estará compuesta por unos y ceros (corresponde a los casos que serán omitidos).

1.6.4 PONDERAR CASOS

Cuando ponderemos casos (botón ) lo que estamos haciendo es darle más peso una variable; el efecto será que cuando hagamos algún análisis el resultado estará clasificado de acuerdo a la variable de ponderación.

En la Fig. 13 tenemos el cuadro de diálogo para ponderar, debemos seleccionar una variable con la cual realizaremos este proceso que afectará al resto de las variables. Para quitar la ponderación sólo debemos presionar en “No ponderar casos”, y después dar clic en Aceptar.

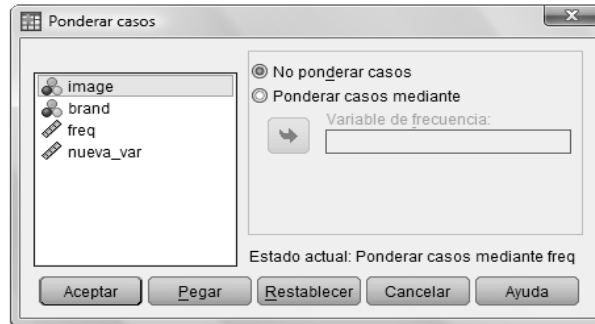


Fig. 13. Ponderar casos

1.7 FUNDIR ARCHIVOS

Mediante esta operación podemos unir los datos de distintos archivos de SPSS, ya sea que tengamos abierto un archivo .sav o también podemos indicarle la ruta en donde se encuentre el otro archivo.

Esta operación maneja dos formas para hacer la fundición, las cuales son:


- **Añadir casos:** Usará un emparejamiento con los casos existentes para añadir los nuevos, de no encontrar coincidencias entre los archivos, simplemente agregará los casos de la variable en cuestión después del último caso.
- **Añadir variables:** Se trata de unir tal cual los archivos, también se puede usar una variable de emparejamiento, ésta debe de estar en ambos archivos.

Cabe resaltar que en ambos casos tenemos que poner atención de cuáles son las variables que seleccionemos, ya que el resto serán eliminadas, y el archivo que sufrirá los cambios será en el que se abra el cuadro de diálogo de “Fundir archivo”.

1.8 SALVANDO NUESTRO TRABAJO

Algo muy importante, que no debemos olvidar, es salvar nuestro trabajo, de hecho esto deberíamos hacerlo desde que insertamos el primer dato.

Ya sabemos los tipos de extensiones de archivo que trabaja SPSS, por lo que guardarlos es cuestión sencilla, podemos hacerlo:

- Usando el botón . Si lo presionamos y no habíamos guardado los datos nos llevará al cuadro de diálogo “Guardar datos como (ver Fig. 14)”, o si ya teníamos guardado nuestro archivo entonces actualiza este archivo con los cambios hechos recientemente.
- Mediante Archivo → Guardar. Cuenta con tres modalidades, las que explicamos antes, y otra más que dice “Guardar todos los datos” que consiste en guardar el archivo tal cual está.

Dentro de la opción de “Guardar datos como” tenemos la posibilidad de guardar ciertas variables del archivo, como vemos en la Fig. 14, presionando el botón “Variables” tendremos otra ventana en la que debemos escoger qué variables formarán parte de nuestro archivo.

Una vez que sepamos cómo vamos a guardar nuestros datos así como la ubicación, debemos nombrar al archivo y dar clic en Aceptar.

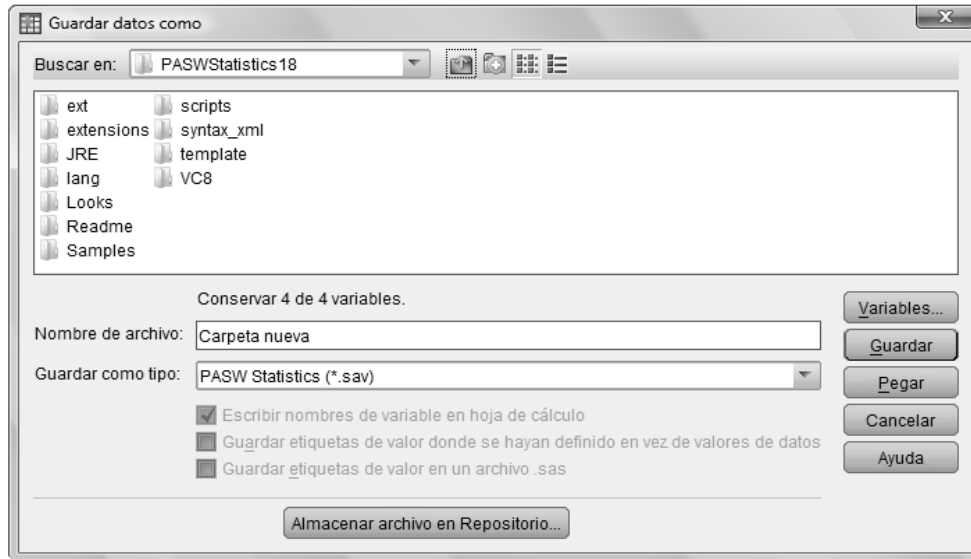


Fig. 14. Guardar datos como

1.9 EJEMPLO GENERAL

Ahora que tenemos una visión de las funciones básicas del programa, realicemos una práctica general para reforzar lo que hemos aprendido hasta ahora, y por qué no aprovechar en ver otras cosas.

Ingresemos a la página web del INEGI, luego en la parte de estadística en la sección de temas demos clic en “Ciencia y tecnología”, teniendo desplegada la categoría de e-Educación daremos clic en “egresados-1969-2001-nacional”, el cual contiene el año de ingreso, la matrícula y el número de egresados; por último en la parte inferior izquierda de la página tenemos un botón que nos ayudará a “Exportar” los datos a un archivo de Excel o Word; para este caso los guardaremos en un archivo de Excel.

Una vez que tengamos nuestros datos, los abriremos en SPSS con el menú Archivo → “Leer datos de texto” y en el tipo de archivo seleccionemos el tipo Excel y luego indiquemos la ubicación, después será necesario seguir el asistente.

Algo muy importante que debemos hacer antes del paso anterior es limpiar la hoja de los datos, de modo que contenga únicamente los nombres de las variables y los datos (sin espacios, y con el formato de número usando la combinación de teclas Ctrl+1), tal vez parezca aburrido pero verás que ocurre cuando los pasemos a SPSS.

Antes de seguir guardemos nuestro archivo, con la extensión “.sav”. Ahora ya tenemos las 4 variables que mencionamos al principio de la actividad, pero... ¿no notas algo extraño?, veamos:

1. Las variables conservan el nombre con respecto al archivo de Excel.
2. La variable primeringreso tenía un espacio que el programa quitó.
3. Por último, revisa qué pasó con la vista de variables, ¡no olvides que son las características de las variables!

Gracias a la limpieza y formato que le dimos al archivo, es más fácil darle el resto de los atributos a las variables, estos quedarán como se muestra en la tabla 1. ¿Alguna duda sobre el cuadro anterior? Pues bien, la primer variable no fue de tipo fecha debido a que SPSS no contempla formatos “solos”, sino combinaciones de meses, años, días, minutos, etc.; por tal motivo es más práctico darle el atributo numérico.

Nombre	Ti-po	Anchura	Deci-males	Etiqueta	Valo-res	Perdi-dos	Colu-mnas	Aline-ación	Me-dida	Rol
Periodo	N	11	0	Año	Nin.	Nin.	11	D	Ordin-al	Nin.
Primeringreso	N	11	0	Primer ingreso	Nin.	Nin.	11	D	Nomi-nal	Nin.
Matrícula	N	11	0	Matrícula	Nin.	Nin.	11	D	Nomi-nal	Nin.
Egresados	N	11	0	Número de egresados	Nin.	Nin.	11	D	Nomi-nal	Nin.

NOTAS: N=Numérico; Nin.=Ninguna, D=Derecha

Tabla 1. Atributos

Todas las variables, excepto “Periodo”, son Nominales ya que no siguen un orden lógico, sólo miden algo. Por último, el rol no lo definiremos.

Antes de comenzar con los cálculos definiremos una hipótesis acerca de la información a utilizar, es decir, ¿qué impresión tenemos acerca de las estadísticas de los egresados de carrera que tengan que ver con ciencias y tecnología? En primera instancia, diremos que de acuerdo al creciente uso de las tecnologías el número de estudiantes en licenciaturas involucradas con Ciencia y tecnología debieron haber aumentado, sin embargo, el número de egresos no tiene que ser proporcional, ya que se involucran factores externos como: el interés del alumno, problemas socioeconómicos, ocurrencia de algún fenómeno natural que afecte al alumno, cambio de carrera, entre otros.

El análisis más sencillo y usado, es el de la participación, es decir, en qué porcentaje contribuyó la variable X en el estudio Y. Ahora obtengamos qué porcentaje representa el número de egresados con respecto al de primer ingreso; y qué ocurre con respecto a la matrícula y egresados; para calcular estas variables veamos la Fig. 7, e ingresemos la expresión “Egresados/Primeringreso”, a esta nueva variable la nombraremos “Egre-ingre”.

No olvides darle los atributos pertinentes, usando el botón de “Tipo y etiqueta”. El proceso es el mismo para el cálculo del porcentaje de egresados con respecto a la matrícula.

Los datos están por año, desde 1969 hasta 2001, ahora necesitamos ubicar a los periodos 69-79 como 1, 80-89 con un 2 y, 90-01 como 3, para esto usaremos la “recodificación en distintas variables”; ya que estemos en el cuadro de diálogo de la Fig. 7, tendremos que definir los nuevos valores con la opción “Valores antiguos a nuevo”, (en la Fig.15, ya están señalados los rangos de la recodificación).

El siguiente paso es calcular la diferencia porcentual de la matrícula con respecto al ingreso, pero únicamente de los 80’s. ¿Cómo lo haríamos?

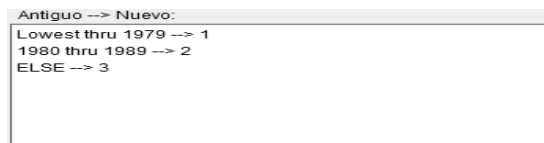


Fig. 15. Recodificación

Existen dos caminos:

1. Dado que ya tenemos recodificada la variable Periodo, podríamos usar un filtrado sobre esa variable y descartar los valores 1 y 3; y después hacer el cálculo.
2. Otra forma es usando “Si la opción...”, de la operación calcular variable.

Usemos primero el filtro; y activemos la opción de “Si satisface la condición”; una vez ahí debemos de introducir la expresión “per_agrup=2”.

Luego de lo anterior hagamos el cálculo de la diferencia ((Matrícula-Ingreso)/Matrícula); y... ¿vualá?, el cálculo se hizo para todos los datos de la variable, esto se debe a que el filtrado no aplica para este tipo de operación.

En cuanto a la segunda forma de obtener la diferencia, condicionaremos, como ya dijimos, con “Si la opción...”, luego debemos activar “Incluir si el caso satisface la condición” y poner la condición “per_agrup=2”.

Para recordar la función de “Ordenar casos”, dejemos este archivo ordenado de acuerdo a la variable egresados, tomando en cuenta que el primer valor será el número mayor.

No en todos los casos podemos aplicar todas las herramientas de SPSS, ya que dependerá del análisis deseado y de la información con la que contemos.

Todos los cálculos hechos anteriormente fueron sencillos, pero no los hemos interpretado, que es lo más importante dentro de un análisis. En resumen tenemos que:

1. Para el año 1991, de 15,793 alumnos de primer ingreso se graduaron 5031, es decir, un 32%.
2. La matrícula del año 1977 fue de 17,777, con 132 egresados, que representan un 7%.
3. En el año de 1987 hubo una diferencia de ingreso con respecto a la matrícula de 72%, lo que significa que el número de alumnos de primer ingreso fue mucho menos que lo que se esperaba.
4. Los años 2001, 2000, 1999; representan los años en los que hubo mayor número de egresados.

La conclusión general de este estudio es que el número de personas que egresan de una licenciatura de “Ciencia y tecnología” ha ido en aumento debido al creciente uso de nuevas tecnologías, sin embargo, a pesar de que la matrícula fue aumentada no tuvo una relación proporcional ni con los ingresos ni egresos de los estudiantes, ¿a qué podemos atribuir estas características? en un primer esbozo podemos decir que se puede deber a la condición socioeconómica del estudiante, a la deserción escolar debido a problemas familiares, o simplemente la falta de interés que sucede por la mala elección de una licenciatura.

Dicha respuesta no la tenemos a ciencia cierta, sin embargo mediante otro estudio enfocado a las causas podremos responder e incluso mejorar las conclusiones previamente dadas.

Como te habrás dado cuenta no exploramos todo lo visto, pero, ¡no os preocupéis!, ya que en el transcurso iremos ocupando parte de lo visto aquí, ya que el objetivo fue familiarizarnos con el entorno de SPSS.

Capítulo 2. ESTADÍSTICA DESCRIPTIVA

La estadística es una ciencia que demuestra que si mi vecino tiene dos coches y yo ninguno, los dos tenemos uno.

George Bernard Shaw (1856-1950)

OBJETIVO: En este capítulo conoceremos la forma de obtener estadísticos descriptivos, además de algunos tipos de gráficos, para describir a nuestros datos a partir de éstos.

La estadística es una de las principales herramientas para tomar decisiones, hacer inferencias acerca del comportamiento de los datos, para la realización de proyecciones, en fin, tiene una amplia utilidad dentro del sector educativo, empresarial, social, entre otras áreas. En general, la estadística se encarga de estudiar fenómenos aleatorios, lo que hace muy amplio su campo de aplicación.

Una de las vertientes de la estadística, es la descriptiva. Con ella podemos saber cómo es la población que se está estudiando, ya que tal como su nombre lo dice, la describe usando medidas de tendencia central, no central, de dispersión y de forma.

La base para la estadística descriptiva son las grandes cantidades de datos, por lo que es de gran utilidad resumirlos mediante su distribución en clases o categorías y definir los elementos que pertenecen a cada una de ellas, cada una de éstas categorías lleva por nombre frecuencia de clase y a la tabulación de éstas le llamaremos: distribución de frecuencias.

Por lo que a SPSS, la mayoría de las pruebas estadísticas se encuentran en el menú Analizar.

2.1 MEDIDAS DE TENDENCIA CENTRAL

Las medidas de tendencia central nos dirán cómo se comportan nuestros datos con respecto a una medida central, dentro de las cuales tenemos:

- **Media:** Mejor conocida como “valor esperado”, da como resultado el promedio aritmético de los valores de la variable que estemos analizando. La obtenemos mediante la suma de todas las observaciones entre el número total de datos. Cabe resaltar que dado que toma todos los valores nuestra media puede verse afectada por la existencia de valores extremos, es decir, aquellos que sean o muy altos o muy bajos.

Su fórmula está dada por:

$$\mu = E(x) = \sum_x x f(x) = \sum_{i=1}^n f_i \left(\frac{1}{n} \right) = \frac{\sum_{i=1}^n x_i}{n} \quad \text{si } x \text{ es discreta}$$

$$\mu = E(x) = \int_{-\infty}^{\infty} x f(x) d(x) \quad \text{si } x \text{ es continua}$$

El valor de x corresponde a cada uno de los valores de la variable analizada. Diremos que se trata de una variable discreta cuando tome valores enteros (por ejemplo: la edad, los meses, nivel de preferencias, el número de prendas compradas), y continua cuando tome valores reales (por ejemplo: el peso, los kilogramos comprados de un producto, los precios del artículo a).

- **Mediana:** Se refiere al valor que se encontrará a la mitad de nuestros datos una vez que estén ordenados ascendentemente, de modo que la mitad de los valores son menores a éste y la otra mitad son mayores. Si el número de datos es impar el valor de en medio es la mediana; sin embargo, si el número de datos es par, entonces se obtiene el promedio aritmético de las observaciones que se encuentren a la mitad del conjunto. A diferencia de la media, la mediana no tiene el problema de que afecte para su cálculo la existencia de valores extremos que afecten su valor.
- **Moda:** Se determina con el valor que ocurra con mayor frecuencia dentro del conjunto de datos. Si el conjunto de datos es muy pequeño, entonces la moda puede no ser muy clara, o si en éste se presenta más de una valor con igual número de repeticiones se dice que tiene más de una moda.

Para que obtengamos las medidas de tendencia central a partir del programa, debemos dar clic en: Analizar → "Estadísticos descriptivos" → Frecuencias. Una vez que estemos en el cuadro de diálogo de la Fig. 16, bastará con que coloquemos la variable sobre la cual queremos los Estadísticos, Gráficos, etc.

En lo que respecta a este capítulo, nos enfocaremos a los estadísticos (Ver Fig. 17) que ofrece el cuadro de diálogo "Frecuencias"; para elegir qué medidas deseamos obtener debemos dar clic en el botón "Estadísticos" y poner las marcas de verificación en la medida de tendencia central deseada, en este caso nos ubicaremos en la sección "Tendencia central", después daremos clic en continuar.

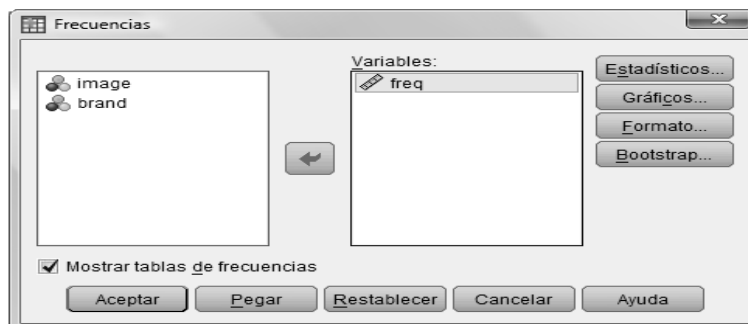


Fig. 16. Frecuencias



Fig. 17. Estadísticos

Si dejamos activo “Mostrar tablas de frecuencias”; como resultado tendremos dos tablas en el Visor de Resultados, una será la de los estadísticos y otra correspondiente a las frecuencias. Cabe mencionar que la tabla de frecuencias es construida a partir de la ocurrencia de cada dato, de modo que para obtener las clases tendríamos que realizar una recodificación.

2.2 MEDIDAS DE DISPERSIÓN

Esta clase de medidas reflejan qué tan dispersos están nuestros datos con respecto al valor central, las más conocidas son:

- Varianza: Es el promedio con el que los datos varían, es decir, qué tan concentrados están los valores alrededor de la media; entre más concentrados estén más pequeña será la varianza, ya que es la distancia entre los datos con respecto a la media.

La fórmula para calcular la varianza es la siguiente:

$$\sigma^2 = E[(x - \mu)^2] = \sum (x - \mu)^2 f(x) = \frac{\sum (x - \mu)^2}{n} \quad \text{si } x \text{ es discreta}$$

$$\sigma^2 = E[(x - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) \quad \text{si } x \text{ es continua}$$

Al igual que en la media, la elección de la fórmula, para el cálculo de la varianza, dependerá de si la variable es discreta o continua.

- Desviación estándar: Es el grado en que las puntuaciones de las variables se alejan; y resulta de la raíz cuadrada positiva de la varianza.
- Rango: Es la diferencia entre el mayor y el menor número de todo el conjunto de datos.

Las medidas de dispersión con las que SPSS cuenta las encontramos en la sección “Dispersión” de la Fig. 17. Para obtenerlas debemos colocar la marca de verificación en la medida de dispersión que necesitemos.

Otra forma para que obtengamos los estadísticos de dispersión es con: Analizar → “Estadísticos descriptivos” → Descriptivos → Opciones.

2.3 MEDIDAS DE FORMA

Antes de que comencemos a conocer cuáles son las medidas de forma, debemos introducirnos al cálculo de los momentos. Existen dos tipos de momentos, los que son alrededor de la media y alrededor de cero, para el caso específico de las medidas de forma debemos usar los momentos alrededor de la media, los cuales son:

$$\text{Primer momento: } \mu_1 = E(x - \mu)^1 = E(x) - \mu = \mu - \mu = 0$$

$$\text{Segundo momento: } \mu_2 = E(x - \mu)^2 = E(x^2) - 2\mu E(x) + \mu^2$$

$$\mu_2 = E(x^2) - \mu^2 = E(x^2) - (E(x))^2 = \mu_2' - \mu^2$$

$$\text{Tercer momento: } \mu_3 = E(x - \mu)^3 = E(x^3) - 3\mu E(x^2) + 3\mu^2 E(x) - \mu^3 = \mu_3' - 3\mu\mu_2' + 2\mu^3$$

Cuarto momento: $\mu_4 = E(x - \mu)^4 = E(x^4) - 4\mu E(x^3) + 6\mu^2 E(x^2) - 4\mu^3 E(x) + \mu^4$

$$\mu_4 = \mu_4' - 4\mu\mu_3' + 6\mu^2\mu_2' - 3\mu^4$$

Cabe mencionar que hay r-momentos, pero únicamente utilizaremos hasta el cuarto momento para el cálculo de las medidas de forma.

Las medidas de forma nos dan una idea de la distribución que siguen los datos, éstas son:

- **Curtosis:** Con esta medida podemos determinar el grado de concentración de los valores en la región central. Gráficamente (ver Fig. 18) si tiene un pico alto, entonces decimos que es leptocúrtica. Si está aplastada, entonces es platicúrtica. Finalmente, si tiene la forma de una curva normal, entonces es mesocúrtica. La fórmula para calcular el coeficiente de curtosis está dada por:

$$\alpha_4 = \frac{\mu_4}{\mu_2^2}$$

En caso de no contar con la gráfica, podemos utilizar el valor de α_4 : si es menor que 3, entonces es platicúrtica; si es mayor que 3, entonces es leptocúrtica; y si es igual a 3, entonces diremos que es mesocúrtica.

- **Asimetría:** Nos permite identificar si los datos se distribuyen de forma uniforme con respecto a la media (ver Fig. 18). Si la gráfica está sesgada hacia la derecha entonces tiene sesgo positivo y se dice que es asimétrica positiva; si el sesgo es hacia la izquierda entonces es asimétrica negativa. La fórmula para obtener el coeficiente de asimetría es la siguiente:

$$\alpha_3 = \frac{\mu_3}{\mu_2^{3/2}}$$

Si el valor de μ_3 es mayor que 0 entonces es asimétrica positiva, si es menor que 0 entonces es asimétrica negativa, y si es igual a 0 entonces es simétrica.

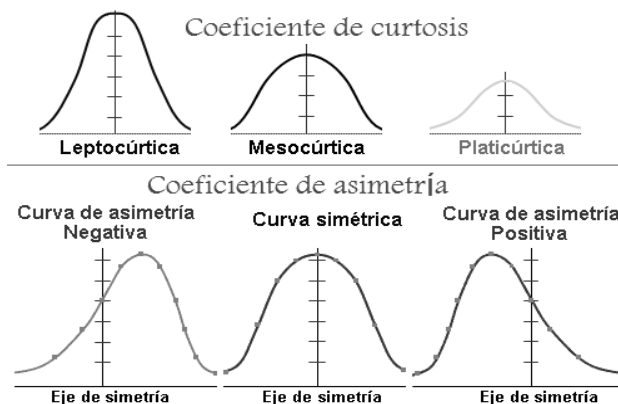


Fig. 18. Coeficientes

Para obtener estas medidas bastará con colocar las marcas de verificación correspondientes en el cuadro de diálogo de la Fig. 17.

2.4 MEDIDAS DE TENDENCIAS NO CENTRAL O DE POSICIÓN

Las medidas de tendencia no central son las correspondientes a los:

- Cuartiles: Son medidas de posición que dividen el conjunto de datos en cuatro grupos, con igual número de términos. Una de las aplicaciones es en la determinación de estratos o grupos, tales como, los socio-económicos, monetarios, niveles de satisfacción. Equivalen a los percentiles 25(cuartil 1), 50 (cuartil 2, también coincide con la mediana) y 75(cuartil 3).
- Deciles: Estas medidas de posición dividen el conjunto en diez partes, equivalen a los percentiles múltiplos de 10.
- Percentiles: Finalmente estas medidas dividen en cien al conjunto de datos, por lo que acumulan un determinado porcentaje de casos.

La obtención de estas medidas con SPSS es la misma que usamos para las medidas de tendencia central, pero ahora nos ubicaremos en la sección “Valores percentiles”.

2.5 GRÁFICOS

Además de las medidas estadísticas, los gráficos forman una parte importante para la descripción de una población. SPSS ofrece varios tipos de gráficos, los cuales están agrupados en el menú de Gráficos.

A continuación hablaremos de los gráficos más usados dentro de la estadística descriptiva, localizados en Gráficos→ “Cuadros de diálogo antiguo”.

2.5.1 GRÁFICO DE DISPERSIÓN SIMPLE

Los gráficos de dispersión representan el grado de relación que existe entre dos variables cuantitativas, el más utilizado es el de dispersión simple; sin embargo, SPSS maneja cinco tipos de gráficos de dispersión: Dispersión simple, Dispersión matricial, Puntos simple, Dispersión superpuesta, Dispersión 3-D.

Para la configuración del gráfico de dispersión simple tenemos que especificar las variables correspondientes al eje X y la al del eje Y.

2.5.2 HISTOGRAMA

El Histograma es un gráfico que está conformado por barras a lo largo de una escala de intervalos iguales, la altura de cada barra corresponde a la ocurrencia de valores que se encuentra dentro de cada intervalo. Este tipo de gráfico nos muestra la forma, la dispersión de la distribución, además de ayudarnos a identificar si los datos están normalmente distribuidos.

Su obtención la podemos hacer mediante el cuadro de diálogo de los Estadísticos del menú de Frecuencias, una vez ahí tenemos que activar la opción “Histograma”. Adicionalmente podemos adjuntar la curva de una distribución normal para comparar ambos gráficos.

También podemos obtener el histograma en el menú de Gráficos requerimos la variable sobre la cual haremos el histograma además de especificar si queremos que nos muestre la curva normal.

2.5.3 GRÁFICO DE BARRAS

Los gráficos de barras son de los más populares y muestran la frecuencia de cada valor o categoría distinta con una barra diferente. SPSS cuenta con los gráficos de barras simple, agrupado y apilado, los cuales explicaremos a continuación.

El gráfico de barras agrupado es el indicado cuando queremos clasificar una variable mediante categorías; de este modo visualizaremos comparaciones de una variable con base en otra; por ejemplo, si necesitamos

saber cómo es la aceptación de los programas de televisión de acuerdo al tipo de programa y el género del televidente, con este gráfico podemos obtener ese contraste agrupado por género y programa de televisión.

Por lo que se refiere al gráfico de barras apilado, éste es parecido al gráfico de barras agrupado, pero ahora cada barra representa una categoría y estará compuesta o bien por una variable o por otra categoría; retomando el ejemplo anterior las barras representarán el género y en cada una de ellas visualizaremos los tipos de programas.

Para configurar cualquiera de los gráficos de barras debemos definir qué representarán las barras, es decir, si se trata del número de casos, porcentaje de casos, porcentaje acumulado. Además de establecer la variable que definirá al eje de las categorías. Si se trata del gráfico de barras agrupado o apilado debemos especificar la variable que defina los grupos o pilas respectivamente, una vez establecido bastará con dar clic en el botón Aceptar.

2.5.4 GRÁFICO DE CAJA

Con este tipo de gráfico veremos la aglomeración de los datos, para su construcción se ocupan los cuartiles o sus equivalentes en percentiles.

Otro aspecto más que nos permite identificar este tipo de gráfico es si dentro de nuestro conjunto de datos hay “datos atípicos”, los cuales estarán muy dispersos de las medidas de tendencia central; algunos analistas consideran que deben ser eliminados, sin embargo, de hacerse podría causar que surgieran otros más, por lo que una buena opción es sustituir el valor que tenga por el de la media, o decidir si conviene obtener otros datos.

Los diagramas de caja con lo que cuenta SPSS, los cuales son: de caja simple o agrupado; estos últimos se manejan de la misma forma que los de barras agrupados. Para configurar estos gráficos debemos definir la variable que representa la caja y el eje de categorías.

Ahora te puedes preguntar, cuándo debemos usar la mediana o un gráfico de sectores, así que a modo de acordeón tenemos la tabla 2, en la que de acuerdo al tipo de variable se especifica qué medida (de tendencia central o no, o de dispersión) y gráfico es el más adecuado.

Tipo de variable	Medida	Gráfico
Nominal	Moda	De barras o de sectores
Ordinal	Mediana, máximo, mínimo, y rango.	De sectores
Intervalo	Media, desviación estándar, máximo, mínimo, coeficiente de asimetría y de curtosis.	Histograma

Tabla 2. Acordeón

2.6 EJEMPLO GENERAL

Dentro de esta unidad vimos lo relativo a cómo obtener estadísticos y gráficos descriptivos, con el siguiente ejemplo exploraremos la forma de analizar lo obtenido con las medidas de estadística descriptiva.

En el buscador tecleemos BIE (Banco de Información Económica), estando ahí demos clic en “Cobertura temática” → “Comunicaciones y transportes” → “Servicio telefónico”; después presionaremos el icono



, el cual nos servirá para hacer la exportación masiva, presionaremos el botón de “Todo” y luego en “Consulta previa”, luego de esto, en “Exportación masiva”. Enseguida guardaremos un archivo de consulta web de Excel, después será necesario abrir el archivo y que al momento de hacerlo nuestro equipo de cómputo cuente con conexión a internet, nos aparecerá la ventana de la Fig. 19.

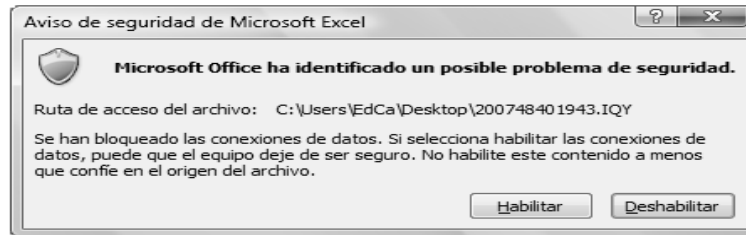


Fig. 19. Seguridad Microsoft Excel

Una vez que habilitemos los datos serán exportados y podremos manipularlos como cualquier otra hoja de Excel, eso sí, debemos de guardar este archivo. En total tendremos 17 variables.

Quizá fue un poco tedioso el proceso para obtener los datos, pero es sólo para la primera vez, la siguiente es realmente rápida y sencilla.

Con estos datos y con las herramientas del software que vimos en esta unidad, analizaremos cómo se comportan los ingresos del sector telefónico, servicio de internet. Larga distancia y celular; desde la perspectiva de los ingresos por tipo de servicio, número de clientes, y minutos de uso.

Al igual que en el ejemplo del capítulo 1, planteemos una hipótesis acerca de los datos a analizar. Diremos que el servicio telefónico para larga distancia por ser un servicio más costoso que el de llamadas locales tiene un mayor ingreso, esto sin dejar de lado que lo relativo a servicios de red ha tenido una mayor influencia a partir del año 2000.

Antes de usar los datos modificaremos el archivo de la forma siguiente:

1. En una hoja nueva tendremos las variables Periodo, Larga distancia internacional, Larga distancia nacional y Servicio local. Será necesario que descombinemos la celda de Periodo, y quitar las celdas que describen la unidad de medida de la variable.
2. En otra hoja colocaremos únicamente las variables: Interconexión, Redes corporativas, Internet y Cuentas de acceso a internet.
3. Crearemos un archivo .txt que tenga las variables: Total de líneas en servicio y Total de clientes celulares. Nombraremos al archivo como "Totales". Antes de que cerremos el archivo, cambiemos los espacios que hay en los nombres de las variables por "_".
4. En un bloc de notas copiaremos las variables Periodo, Nacional e Internacional; guardaremos el archivo con el nombre "Consumoxminuto". No olvidemos verificar que cada dato (incluyendo los títulos), estén separados por un espacio, o para hacerlo más presentable usemos la tecla del tabulador como un separador.

Luego de lo anterior, abriremos el SPSS y exportaremos la hoja de Excel que contenga las variables Periodo, Larga distancia internacional, Larga distancia nacional y Servicio local; este archivo lo nombraremos "comunicaciones".

En la hoja original en la que están todos los datos tenemos la descripción de los símbolos usados, recordemos que los ingresos se refieren a millones de pesos.

Tomando en cuenta la tabla en la que se especifica el tipo de descriptivo, obtendremos: Media, desviación estándar, máximo, mínimo, coeficiente de asimetría y de curtosis; y el histograma para las variables Larga distancia internacional, Larga distancia nacional y Servicio local.

En el ejercicio del capítulo anterior no comentamos cómo se va formando la hoja del visor de resultados, éste es alimentado por la sintaxis de SPSS, tablas, gráficos y errores. Cada uno de ellos es posible editarlo dando doble clic sobre ellos, de tal forma que quede al gusto personal.

A los gráficos les podemos modificar la letra (color, tamaño, tipo), colores de la gráfica (barras líneas, etc., así como del relleno exterior).

En la Fig. 20 tenemos el editor de gráficos, que nos sirve para cambiar el texto o encabezados, será necesario que seleccionemos el elemento, es decir, que se encuentre rodeado por una línea amarilla y luego dar doble clic, de este modo veremos el cuadro de diálogo de la Fig. 21, en donde vienen las propiedades del elemento seleccionado.

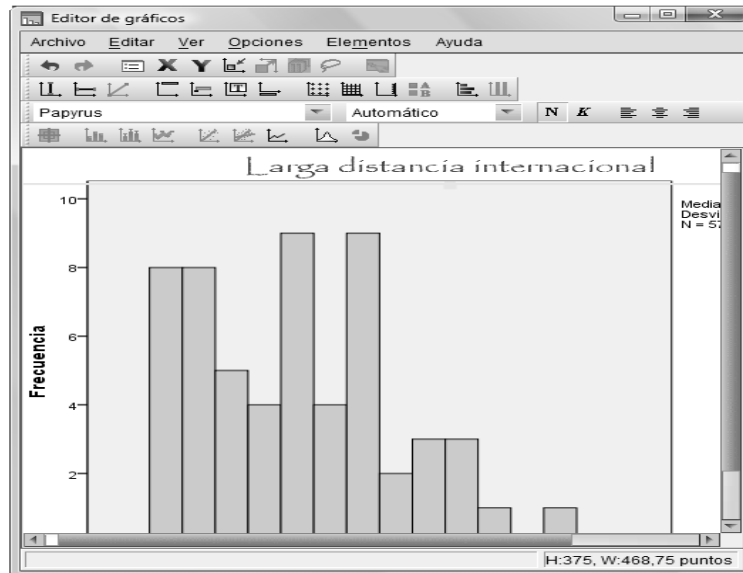


Fig. 20. Editor de gráficos

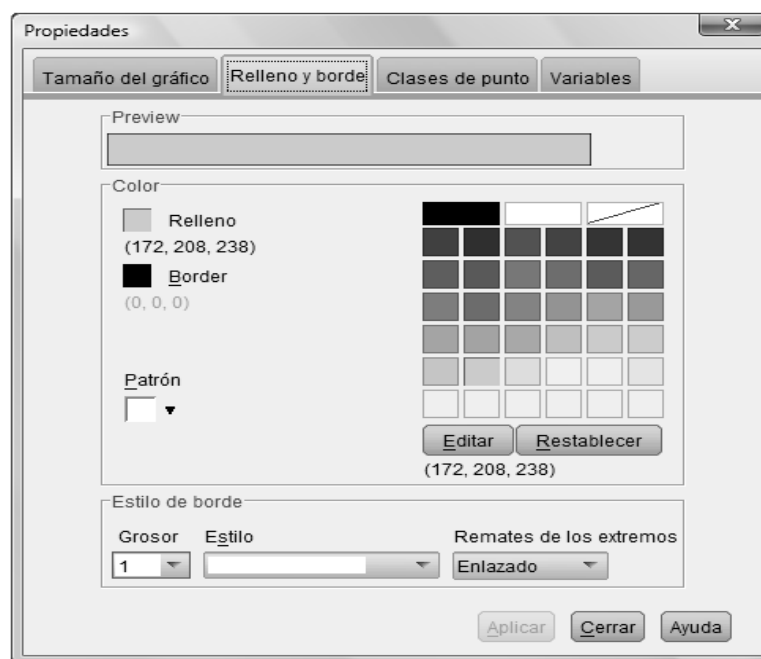


Fig. 21. Propiedades del gráfico

En el caso de las tablas es el mismo proceso, es posible rotar las columnas con las filas, cambiar el color, tamaño y tipo de letra, de esta manera podemos darles un estilo personalizado.

Además de la edición de los resultados, otra de las herramientas que tenemos en el visor de resultados es la exportación de resultados en archivos de diferentes extensiones dependiendo el tipo de resultado. Las extensiones html, pdf, txt, ppt, y doc son válidas para todo tipo de resultados (tablas, sintaxis, gráficas); mientras que las extensiones jpg, png, bmp son las adecuadas para imágenes.

Podemos exportar todo, sólo las tablas o sólo las gráficas; la forma de hacerlo es dando clic derecho ya sea en el área de resultados o dentro del área de información, hecho esto visualizaremos una ventana con la que elegiremos el tipo de archivo, y lo que queremos exportar; sin embargo, para la selección también tenemos la opción de usar el menú Edición → Seleccionar.

Antes de que exportemos, no olvidemos establecer la ruta en la que será guardado nuestro archivo, ya que si no la cambiamos, éste será salvado en la carpeta que crea SPSS. En la Fig. 22 tenemos la ventana que nos indica que la exportación está en proceso.

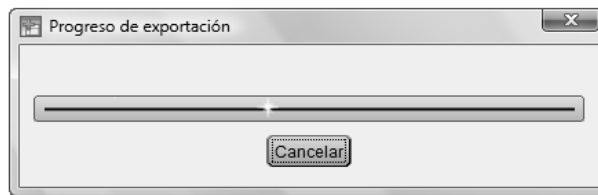


Fig. 22. Exportando...

¡Fabuloso!, ¿no crees?, ahora vayamos a la parte de la interpretación.

La Tabla 3 es el resultado de haber cambiado el estilo y ser exportada a un archivo de extensión .doc. El software reparte el contenido de lo exportado dentro de la página del tipo de archivo seleccionado, por lo que ésta es sólo una parte de la tabla original.

Las tres variables tuvieron 57 casos y ningún valor perdido; dentro de esta primera parte podemos concluir que la media de la variable larga distancia internacional es de 6,934.04.

	N		Media	Desviación típica	Varianza
	Válidos	Perdidos			
Larga distancia internacional	57	0	6934.04	3707.705	13747079.213
Larga distancia nacional	57	0	14686.74	8750.958	76579267.019
Servicio local	57	0	30324.32	16107.863	259463258.291

Tabla 3. Estadísticos 1

La conclusión anterior es la forma incorrecta de reportar un resultado de este tipo, lo más apropiado es:

“En promedio los ingresos trimestrales por llamadas de larga distancia internacional fueron de \$ 6,934.04 millones en el periodo enero de 1996 a enero de 2001.”

“En el periodo de 1996 a 2001 el promedio de variación de los ingresos trimestrales de las llamadas de larga distancia nacional fue de 76, 579, 267.019 unidades al cuadrado.”

“El grado en que varía el ingreso por servicio local de llamadas en el periodo de 1996 a 2001 es de \$16,107.863 millones.”

Con este primer acercamiento podemos darnos cuenta que el Servicio local de llamadas telefónicas es el que tiene un mayor promedio mensual de ingresos, con esto concluiremos que a pesar de que el costo de una llamada de larga distancia nacional o de larga distancia internacional es mayor que el de una de servicio local, su ingreso promedio mensual es inferior, ¿a qué se puede deber? pues bien, podemos decir que es debido a que la población realiza más llamadas locales que de larga distancia, o que debido a que el precio de una llamada de larga distancia es mayor entonces realiza menos llamadas.

Con lo anterior, la primer parte de nuestra hipótesis no fue cierta, ya que los ingresos por el servicio local son mayores al de las largas distancias.

En la Tabla 4 tenemos otra parte de la tabla de estadísticos que obtuvimos, ésta contiene lo relativo a las medidas de forma que vimos en el tema 2.3; de este modo, podemos decir que nuestros datos están concentrados con respecto a la media de una forma leptocúrtica (ver Fig. 18), además no siguen “del todo una distribución normal”, ya que es una curva asimétrica positiva, por tanto su eje de simetría no coincide con el del valor central

	Asimetría	Error típ. de asimetría	Curtosis	Error típ. de curtosis
Larga distancia internacional	.485	.316	-.458	.623
Larga distancia Nacional	.874	.316	.301	.623
Servicio local	.319	.316	-1.040	.623

Tabla 4. Estadísticos 2

El histograma posterior es el correspondiente al de la variable “Servicio local”, a éste le asociamos una curva normal de tal forma que podemos contrastarlo con las barras del histograma y confirmar lo que ya vimos con los coeficientes de asimetría y curtosis, es decir, no sigue una distribución normal

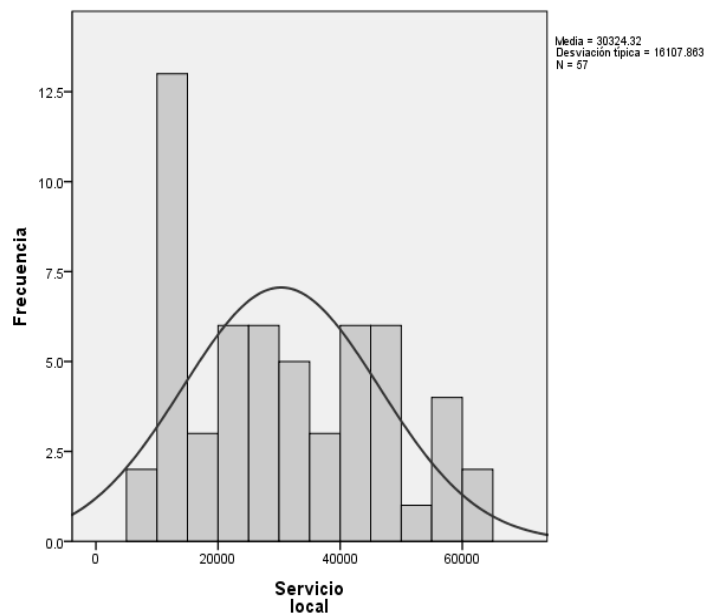


Fig. 23. Histograma Servicio Local

¿Dudas hasta el momento? En el mundo de los negocios no importa la técnica de construcción de nuestros estadísticos, es decir, bien pudimos haber explicado el procedimiento para obtener cada uno de ellos manualmente, pero en este momento estamos resaltando la correcta interpretación de los resultados.

En muchas ocasiones la información está repartida por diversas áreas, y cada una de ellas reportará a un departamento superior. Al principio de la actividad dividimos el archivo original en distintos tipos de archivos, de tal forma que ahora nos encargaremos de unirlos usando la fundición de archivos, haciendo la analogía que cada uno de estos son reportes de otros departamentos.

Antes de hacer la fundición, recordemos que los archivos deben de ser específicamente de extensión **.sav**; para ello debemos de exportar los datos como lo hicimos en la actividad del capítulo anterior, pero ahora para extensiones .txt, .doc; además de la .xls.

Después de que hayamos exportado nuestros archivos a SPSS, seguramente notamos que hacerlo de un archivo de texto tiene sus ventajas, ya que podemos definir cómo van nuestros datos, los atributos que deben tener, o si deseamos una parte o la totalidad de los datos.

Al exportar de un archivo de Excel tenemos la opción de elegir cuál de las hojas que este contiene, ésto nos permite exportar la hoja en la que colocamos las variables Interconexión, Redes corporativas, Internet y Cuentas de acceso a internet.

Otro paso más que debemos hacer es el de dejar completamente listos nuestros datos; primero, en los datos relativos a internet tenemos las variables “redes corporativas” e “Internet” que deben ser numéricas pero que están como cadena, por lo que debemos recodificar en otra variable de la forma siguiente:

- El valor antiguo será “ND”, y para el nuevo activemos “Perdido por el sistema”.
- En la parte final marcaremos “Convertir cadenas numéricas en números”.

El proceso anterior es el mismo para la variable “Internet”, “Total_de_clientes_celulares” (del archivo “Totales”), por lo que bastará agregarlas al mismo cuadro de recodificación, sin olvidar definir la nueva variable para cada una.

Fundiremos el archivo “comunicaciones” con el de “totales” usando “Añadir variables”, excluyendo la variable “Total_de_clientes_celulares”; este proceso es el mismo para fundir el archivo “Internet”, y excluirémos: redescorporativas e internet.

Por último fundamos el archivo “consumoxminuto”, en este caso no usaremos “Añadir casos” porque a pesar de tener una variable en común el efecto que tendría hacer esto es que además de agregar la variable SPSS nos colocaría los valores a partir del último caso. Ahora sí ya tenemos nuestro archivo con 12 variables.

Para continuar tenemos que crear una variable que corresponda sólo al año, esto lo haremos con la función CHAR.SUBSTR ubicada dentro de la categoría “Cadena”, mostrada a continuación:

CHAR.SUBSTR(Periodo,1,4)
 ↑ ↑ ↑
 1 2 3

1. Variable de tipo cadena, sobre la cual se va a sustraer una subcadena.
2. Indica la posición del inicio de la subcadena.
3. Se refiere al tamaño de la subcadena.

El siguiente paso es convertir esta nueva variable de tipo cadena a numérico con la función:

NUMBER(Año,F4)
 ↑ ↑
 1 2

1. Variable de tipo cadena que queremos convertir a numérica.
2. Formato de impresión, este se refiere a la presentación que tendrá la variable en la impresión en pantalla.

Ahora tenemos lista nuestra variable que utilizaremos para segmentar el archivo. Luego de esto obtendremos los estadísticos descriptivos de todas las variables.

¿Cómo interpretaremos la tabla resultante? Pues bien, en esta tabla tenemos los estadísticos descriptivos de cada variable, pero por cada año, es decir, cómo se comportaron nuestros datos por año.

Gracias a lo anterior podemos hacer algunas conclusiones como las siguientes:

- En 1996 hubo un máximo de 657 miles de clientes por el servicio de celular.
- En el año de 1997 en promedio se consumieron 5118.25 millones de minutos por llamadas de larga distancia nacional, y 2,366 millones de minutos por llamadas de larga distancia internacional.
- Para 1998 el ingreso mínimo por interconexión fue de 559 millones de pesos.
- En 1999 el ingreso mínimo por llamadas de larga distancia internacional fue de 2,201 millones de pesos.
- Para 2001 las cuentas de acceso a internet tuvieron un máximo de 845 miles de clientes.
- El año 2002 tuvo en promedio 14,023.5 miles de líneas en el servicio de líneas telefónicas.
- Habría que investigar cuál es la razón por la que no hay registros del número total de clientes por servicio telefónico celular, pero si lo hubo en los años 1996 a 2000.
- El servicio de Internet tuvo un ingreso trimestral promedio de 4,621.75 millones de pesos en el año de 2004.
- El servicio de “Redes corporativas” fue el más representativo en el año 2005 en cuanto a ingresos, teniendo un máximo trimestral de 20,324 millones de pesos.
- Por último, sin desmerecer la información de los años anteriores, para enero de 2010 se tiene registrado un total de 15,811 miles de líneas en servicio, tomando en cuenta líneas con adeudos no mayores a 3 meses.

¡Interesante!, las conclusiones anteriores las seleccionamos al azar, sin embargo, podríamos ahondar más y realizar un análisis más detallado, esto es sin dejar de lado que esta información es sólo la correspondiente a “los resultados relevantes” de Teléfonos de México.

Con esto podemos determinar que en efecto los ingresos por servicio de red sí son representativos, y que si bien en estos servicios estaban disponibles desde antes es en años posteriores al 2000 en los que tomó mayor fuerza, o al menos eso pasó con Teléfonos de México.

Para terminar analicemos el gráfico de caja (ver Fig. 24) que corresponde al total de clientes celulares. ¿Qué importancia tiene este gráfico? A partir de él podemos observar si hay “datos atípicos”, es decir, aquellos datos que no van de acuerdo a lo que mide la variable, por ejemplo, si tuviéramos una variable de nombre Edad y esta tuviera datos superiores a 200 o que sean negativos, esto es totalmente ilógico, debido a que no hay edades que sean menores que cero y tampoco personas que cuenten con una edad incluso superior a 120 años (o al menos son muy pocas).

En resumen, los datos de esta variable están “limpios”, ya que de haber datos atípicos veríamos algún punto fuera de la caja.

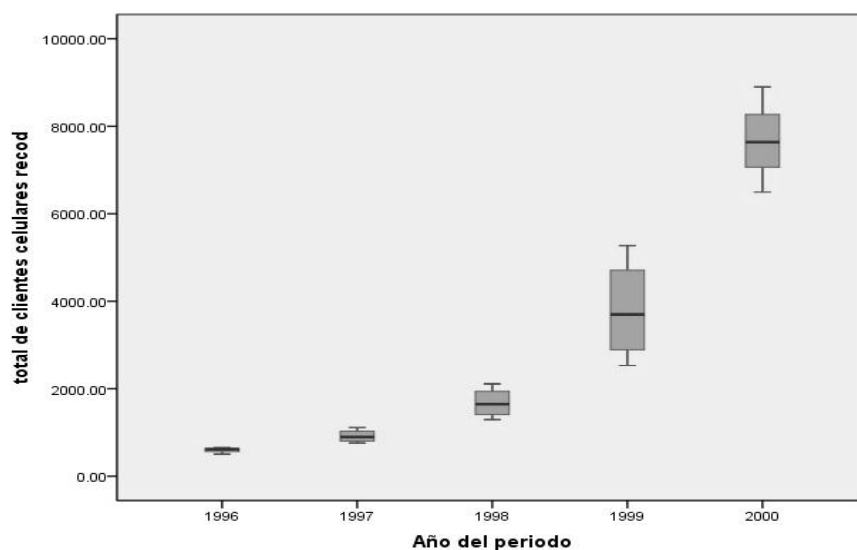



Fig. 24. Diagrama de caja clientes celular

2.7 EJERCICIO COMPLEMENTARIO

En el ejemplo anterior debido a la naturaleza de los datos no fue posible usar todo lo visto en este capítulo, así que toca emplear otro tipo de información. Ingreseemos, en nuestro navegador de internet, a la dirección: <http://www.estadistica.unam.mx/>.

Esta página es el Portal electrónico de Estadística Universitaria de la UNAM, en la que reporta datos relacionados con los alumnos, planes y programas de estudio, Facultades, carreras acreditadas, educación continua, investigación, presencia, servicios nacionales, infraestructura, presupuesto, y los académicos.

En la categoría “SIDEU” daremos clic, luego en el menú elegiremos “Series Históricas de Información Universitaria”, luego un clic en el icono  de “Población Escolar (Sin posgrado)” y empezaremos a descargar los datos en un archivo de Excel.

Cuando abramos el archivo aparecerá el cintillo “Vista protegida”, debemos dar clic en “Habilitar edición”. En la primera hoja tenemos una tabla dinámica que resume la población total por generación, en la segunda hoja está toda la información, la cual contempla las variables: Nivel, Sistema, Plantel, Carrera, Consejo Académico de Área, Generación, Género, Tipo de Ingreso y Población.

Después de haber exportado los datos al software vamos a dar clic en “Variables” y desmarcar “ConsejoAcademicodeÁrea”, tal como vemos en la Fig. 25.

Luego de esto guardaremos el archivo, pero la hoja en la que exportamos se quedará activa, así que será necesario que abramos el archivo que ya habíamos guardado.

El siguiente paso será recodificar en la misma variable a Género, con 0= Mujeres y 1= Hombres, y hacer esa relación en la etiqueta de valor; del igual forma recodificaremos a la variable “TipodeIngreso”, como 0=Reingreso y 1=”Primer Ingreso”.

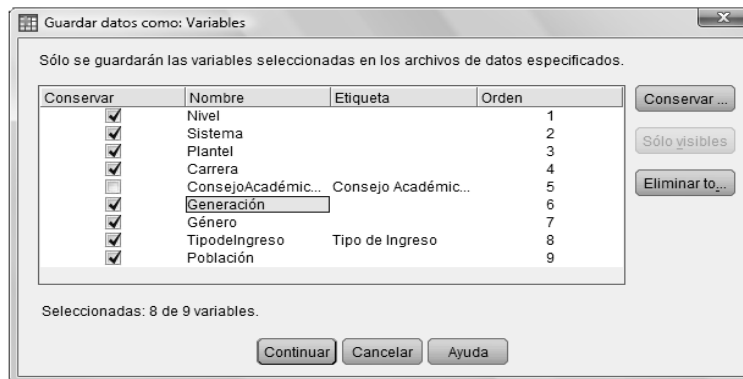


Fig. 25. Variables

La parte anterior es lo inverso a lo que hacen las empresas encargadas de estudios de mercado, ya que ellos manejan variables categóricas, es decir, sólo contienen números que están relacionados con los valores que tiene la misma.

Continuemos con la segmentación del archivo usando las variables “Género” y “TipodeIngreso”, para después hacer un filtrado que nos servirá para excluir los casos relativos a “Propedéutico de la Escuela Nacional de Música”, ya que son datos que no están completos como el resto.

Enseguida obtengamos a partir del menú “Frecuencias” el valor de la mediana, máximo y mínimo (ver tabla 5). Con esto vemos que la población está equilibrada tanto en género como en el Tipo de Ingreso, a modo de ejemplo de conclusión tenemos:

“En promedio hay 214 mujeres de Reingreso y 92 hombres de Primer Ingreso”.

		Población			
		Mujeres		Hombres	
		Reingreso	Primer Ingreso	Reingreso	Primer Ingreso
N	Válidos	4370	4370	4370	4370
	Perdidos	0	0	0	0
Mediana		214.00	75.00	267.50	92.00
Mínimo		0	0	0	0
Máximo		5189	2342	9623	3244

Tabla 5. Frecuencias con segmentación

Lo anterior es en cuanto a la forma de ingreso, ahora veamos cómo están repartidos los niveles de escolaridad de acuerdo al género. La proporción entre los géneros se conserva entre los niveles de escolaridad, dentro de estos, el más representativo es el nivel de Licenciatura (ver Fig. 26).

En el gráfico de barras agrupado por sexo y por Sistema de la Fig. 27; vemos que el sistema escolarizado tiene el mayor número de alumnos, y el Colegio de Ciencias y Humanidades es uno de los sistemas con menor planta estudiantil.

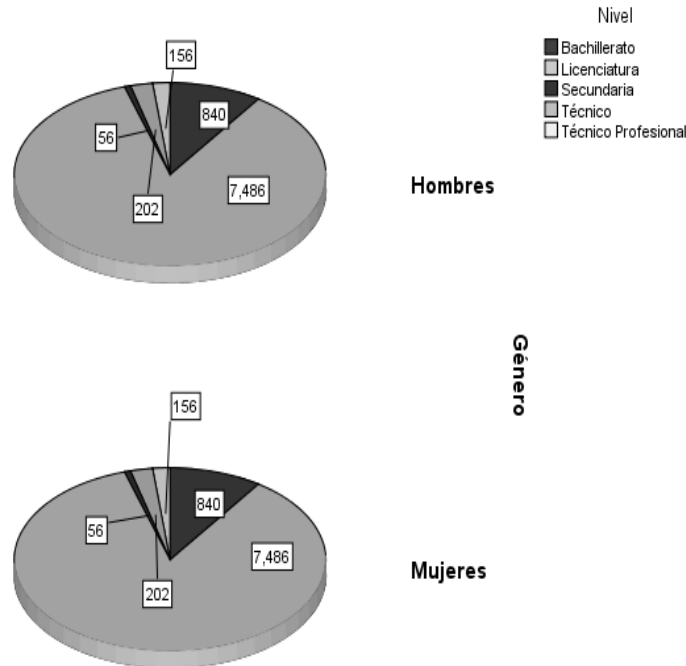


Fig. 26. Nivel vs. Género

Como último paso agrupemos la variable población usando el menú de la “Agrupación visual”, los puntos de corte serán ocupados por “Puntos de corte en media y desviaciones típicas...”, marcaremos los tres cuadros que hay debajo, por último daremos clic en crear etiquetas y ¡listo!

Este tipo de agrupación tiene que ver con la famosa “Regla empírica”, que dice lo siguiente:

- El valor de la media más menos una desviación estándar, comprende aproximadamente el 68.26% de la población (ver Fig. 28).
- La media más menos dos desviaciones estándar comprende aproximadamente el 95.44% de la población (ver Fig. 28).
- El valor de la media más menos tres desviaciones estándar comprende aproximadamente el 99.73% de la población (ver Fig. 28).

Cabe resaltar que esta Regla empírica aplica para datos cuya distribución de probabilidad es parecida a la curva normal, de tal forma que es bastante usada para dar conclusiones como la siguiente:

“El 95% de la población se encuentra entre VALOR1 (el valor de la media menos 2 veces la desviación estándar) y VALOR2 (el valor de la media más 2 veces la desviación estándar)”

En este capítulo vimos los estadísticos descriptivos, los cuales son muy usados en el mundo de las decisiones, que es también el de los negocios.

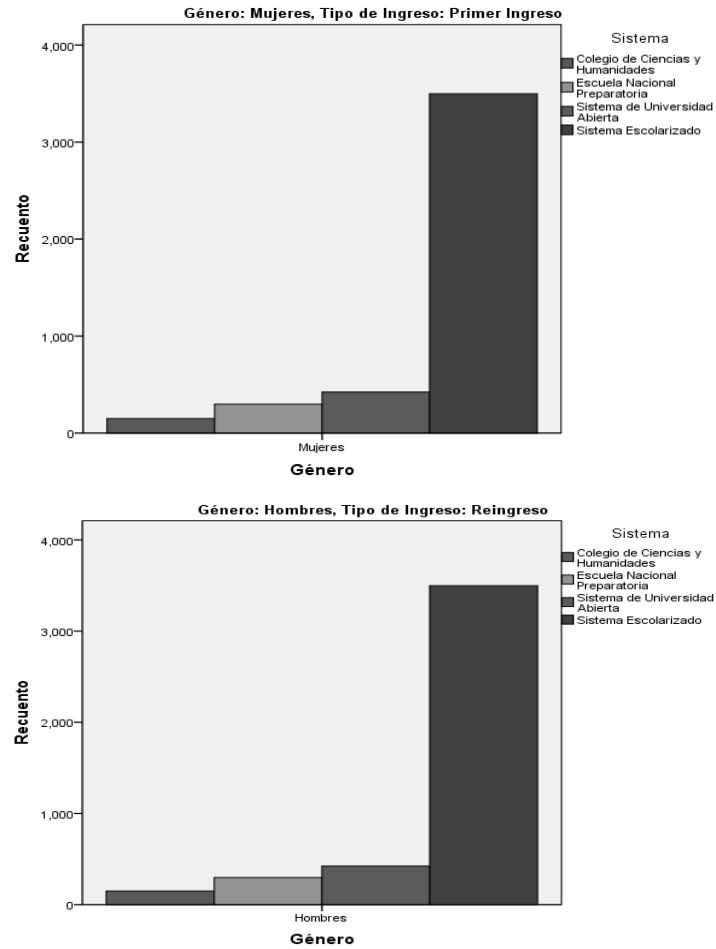


Fig. 27. Sistema vs. Género

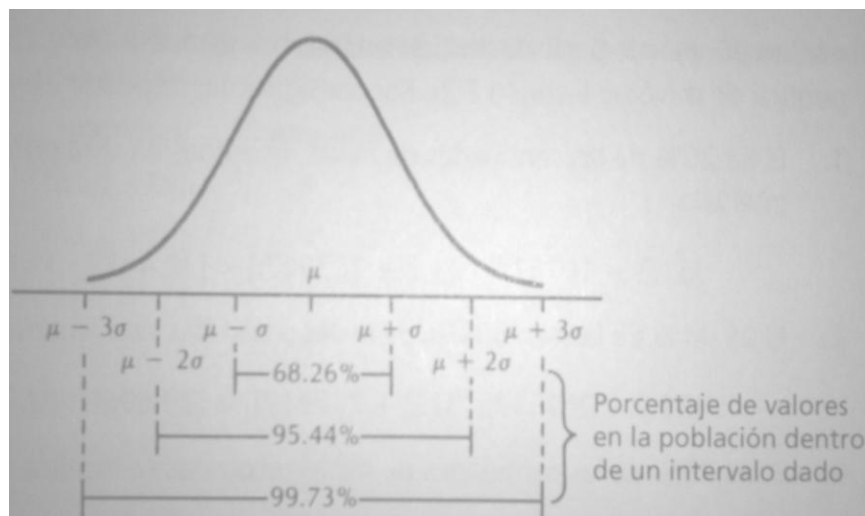


Fig. 28. Porcentajes con respecto a la media y desviación estándar con una población normalmente distribuida. (Fuente: Bowerman, 2007)

2.8 CASO DE ESTUDIO

Como bien mencionamos en la introducción, tenemos la hipótesis de que si la estadística para los alumnos resulta muy teórica y tediosa debido a la realización de análisis como obtener promedios, encontrar valores en tablas, regresiones lineales; entonces el uso de un software estadístico para aplicaciones prácticas reales facilitará el aprendizaje e interpretación de la estadística; para la cual se propone el uso de software estadístico, así como ejemplos reales.

Tomando en cuenta lo anterior se realizó un pequeño experimento en el que se compararon dos grupos. En uno de ellos se utilizó SPSS y en el otro grupo no, además, para ambos se utilizó el ejemplo general del presente capítulo.

La evaluación de la práctica consistió en:

- Tiempo que le tomó al integrante hacer la práctica. Cabe señalar que no hubo un tiempo límite para su resolución.
- Preguntas específicas, correspondientes al análisis y obtención de estadísticos descriptivos, dichas preguntas fueron las mismas para ambos grupos.

Adicionalmente, con el fin de contar con la opinión acerca del manejo de ejemplos reales y de la oportunidad del uso de software estadístico en la enseñanza de la estadística, se realizó una encuesta en ambos grupos.

La encuesta correspondiente al grupo A, correspondiente a las personas que no hicieron uso del software SPSS, fue la siguiente:

<p>ENCUESTA GRUPO A</p> <p>Edad: _____ Género: _____ Carrera cursada: _____</p> <p>1. ¿Tenía algún gusto por la estadística? _____</p> <p>2. ¿Utiliza con frecuencia algún tipo de análisis estadístico? _____</p> <p>3. Si en la carrera cursada tomó algún curso de estadística, ¿Qué tan satisfecho quedó acerca de este en término de lo que aprendió? _____</p> <p>4. Le hubiera interesado que en su experiencia con la estadística se hubiera tomado en cuenta algún software estadístico. _____</p> <p>5. El uso de ejemplos reales facilitó la comprensión del tema visto? _____</p> <p>6. ¿Considera que el tiempo usado para la resolución de la práctica se hubiera minimizado con el uso de algún software de estadística? _____</p>
--

Por lo que se refiere a la encuesta para el grupo B, únicamente cambió la se modificó la pregunta 6, quedando de la forma siguiente:

<p>6. ¿Considera que el uso de SPSS facilitó el aprendizaje del tema visto? _____</p>

Las variables que se manejaron fueron:

- ID: Se trata de una clave que identificará a cada integrante del grupo.
- Género: Es el género de la persona evaluada.
- Método: Definirá si se trata de una persona que usó o no el software.
- Hora de inicio: Se refiere a la hora en que inició la práctica.
- Hora de término: Es la hora en que el integrante terminó sólo la práctica, sin contar la encuesta.
- Tiempo usado: Se trata de la diferencia en minutos de las variables Hora de inicio y Hora de término.
- P1 a P6: Son las variables correspondientes a las calificaciones de las seis preguntas que conformaron la práctica.
- Promedio de calificaciones: Es el promedio de las puntuaciones de las variables P1 a P6.

Muy bien, el análisis medular de esta prueba consistió en saber qué tan bueno resulta o no usar SPSS y ejemplos reales para el aprendizaje e interpretación de la estadística, de modo que se utilizaron pruebas de hipótesis para verificar la hipótesis.

Tomando en cuenta lo anterior, las pruebas estadísticas que se utilizaron para el análisis de los datos fueron:

- Prueba K-S para una muestra. Con ella identificamos si los datos siguen una distribución Normal y así saber si es posible usar una prueba paramétrica o no paramétrica.
- Prueba T para muestras independientes o de la U Mann-Whitney, dependiendo de los resultados de la prueba K-S para una muestra.

Se realizó cada prueba para las variables Tiempo usado y Promedio de calificaciones. En el capítulo 3 veremos los resultados de las pruebas y el análisis de las mismas.

Capítulo 3. PRUEBAS DE HIPÓTESIS

No rechaces a la H_0 , en vez de eso acéptala como parte de tu vida.

Anónimo

Objetivo: En este capítulo revisaremos distintas pruebas de hipótesis, para aceptar o rechazar inferencias que hagamos acerca de una población dada.

Las pruebas de hipótesis son parte de la estadística inferencial, a partir de ellas podemos hacer inferencias acerca de la población³ a partir de una muestra aleatoria⁴; de modo que nos permite aceptar o rechazar, según sea el caso, las conjeturas hechas.

Los elementos principales de un ensayo de hipótesis son:

1. Hipótesis nula (H_0): Está asociada a una comparación de igualdad, es la que se desea comprobar.
2. Hipótesis alterna (H_1): En este caso se establece que el valor del parámetro es mayor o igual, o simplemente que sea diferente, es lo que acepta en caso de que H_0 no sea verdadera.
3. Estadístico de prueba: Es el valor con el que se contrastará la prueba de hipótesis.
4. Valor crítico: Es el valor que toma el estadístico de prueba.
5. Región crítica: Se refiere a la zona en la que están los valores en los que la decisión es rechazar la hipótesis; su área es igual al tamaño del error de tipo I.
6. Regla de decisión o criterios de rechazo: Se formula a partir de la hipótesis nula y alternativa.
7. Nivel de confianza ($1 - \alpha$): Está unido al nivel de significancia, de tal forma que si $\alpha = .05$ entonces se tendrá un nivel de confianza del 95%.
8. Nivel de significación (α): Se refiere a la probabilidad máxima con la que el ensayo puede estar dentro del error tipo I. Los niveles más comunes son .05 y .01.

Una interpretación más clara del nivel de significancia es, para el caso de .05 hay aproximadamente cinco ocasiones en cien en que se rechazaría la hipótesis cuando debería ser aceptada.

9. Error del tipo I y del tipo II: Ocurren cuando:

³ Colección completa de elementos acerca de los cuales se desea información.

⁴ Las observaciones se hacen al azar y cada una de ellas es independiente.

	Aceptar H_0	Rechazar H_0
H_0 verdadera	Decisión correcta	Error del tipo I
H_0 falsa	Error del tipo II	Decisión correcta

Tabla 6. Error tipo I y II

Los tipos de pruebas de hipótesis dependerán de la desigualdad a la que estén sujetas, si se trata de una desigualdad estricta ($<$, $>$), entonces es de una sola cola, por ejemplo:

$$\begin{array}{ll} H_0 : \theta = \theta_0, & H_0 : \theta = \theta_0, \\ H_1 : \theta > \theta_0 & H_1 : \theta < \theta_0 \end{array}$$

Cuando se trata de una prueba bilateral o de dos colas tiene la forma:

$$\begin{array}{l} H_0 : \theta = \theta_0, \\ H_1 : \theta \neq \theta_0 \end{array}$$

En el primer caso la región crítica se localiza a la izquierda o a la derecha, mientras que en la prueba bilateral la región crítica se divide en dos partes que a menudo tiene probabilidades iguales ubicadas en cada cola de la distribución.

3.1 COMPARACIÓN DE MEDIAS

Muchos de los estudios van enfocados a comprobar resultados de tratamientos o comportamientos entre grupos, es aquí en donde entran las pruebas de hipótesis.

Dentro del menú Analizar → “Comparar medias” tenemos el catálogo de pruebas de hipótesis sobre medias.

3.1.1 MEDIAS

Este procedimiento ofrece estadísticos descriptivos tomando en cuenta los grupos que existan dentro de los datos, por ejemplo: mujeres-hombres, tarde-noche.

En la Fig. 29 tenemos el cuadro de diálogo del procedimiento “Medias” así como las “Opciones” que ofrece. En primera instancia colocaremos a la variable cuantitativa (correspondiente a una variable dependiente), para la casilla de la variable independiente necesitaremos una variable categórica.

La media, el número de casos y la desviación, son los estadísticos que SPSS proporciona por default; sin embargo, dentro de “Opciones” podemos seleccionar más estadísticos o incluso quitar los que tenemos en la casilla.

Si utilizamos más de una capa (ver Fig. 29) el resultado estará subdividido en otros más usando otra variable; por ejemplo, en una capa usaremos la variable que se refiera a la región a la que pertenezca una persona, y en una segunda capa tendremos la variable que defina el género.

Los resultados aparecerán de acuerdo a la variable independiente, y por cada capa elegiremos los estadísticos que deseemos.

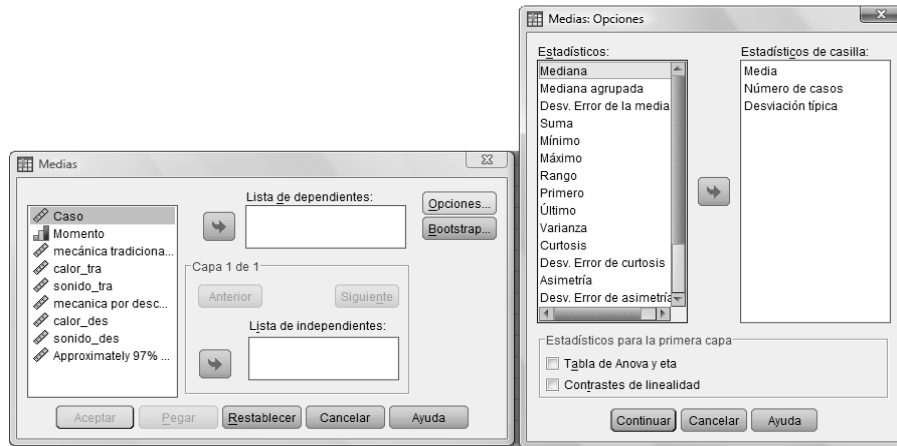


Fig. 29. Medias

3.1.2 PRUEBA T PARA UNA MUESTRA

La prueba T está diseñada para hacer ensayos de hipótesis acerca de la media de una población, bajo los supuestos de que la población de la que obtuvimos la muestra tiene una distribución normal (con media μ y varianza 1), y que el valor de σ no lo conocemos, por lo que usamos a S^5 para estimar su valor.

Veamos, únicamente, cuando se conoce el valor de la desviación estándar y la población de la que se obtuvo la muestra tiene una distribución normal podemos usar el estadístico de prueba:

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

En donde:

\bar{X} : Media de la muestra

μ : Media poblacional

σ / \sqrt{n} : Desviación estándar de la muestra

Sin embargo, cuando se desconoce el valor de la desviación estándar de la población recurriremos a la distribución T, la cual define como variable aleatoria a la siguiente:

$$T = \frac{Z}{\sqrt{X/v}} \quad \text{con } -\infty < T < \infty$$

En donde:

Z: Variable aleatoria con distribución Normal

⁵ S^2 es un estimador de la varianza con distribución chi-cuadrado con (n-1) grados de libertad definido por: $\sum_{i=1}^n (x_i - \bar{x})^2 / (n - 1)$, por lo tanto S es el estimador de la desviación estándar.

X: Variable aleatoria con distribución chi-cuadrada, $(n-1)s^2/\sigma^2$ con $v = n-1$ grados de libertad

v: Grados de libertad, con $v > 0$

Sustituimos z y X en la definición de T:

$$T = \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)s^2/\sigma^2}{(n-1)}}}$$

Eliminando el término n-1, y obteniendo la raíz tenemos:

$$T = \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\frac{s}{\sigma}}$$

Finalmente, aplicando la división:

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

Éste será nuestro estadístico de prueba con (n-1) grados de libertad, quedando nuestra prueba de hipótesis de la manera siguiente:

$$\begin{aligned} H_0 : \quad & \mu = \mu_0, \\ H_1 : \quad & \mu \neq \mu_0 \end{aligned}$$

Por lo que se refiere la regla de decisión, rechazaremos H_0 cuando $T \leq t_{\alpha/2, n-1}$ o cuando $T \geq t_{1-\alpha/2, n-1}$.

Ahora, para definir nuestro valor con el que definiremos nuestra hipótesis nula ingresaremos al cuadro de diálogo de la Fig. 30 y seguiremos los pasos que describiremos a continuación.

En “Valores para contrastar” colocaremos la variable sobre la cual queremos hacer la prueba de hipótesis (por cada variable se hará una prueba T) y en el “Valor de prueba” escribiremos el valor de la media a contrastar.

Dentro de las opciones (ver Fig. 31) podremos especificar el nivel de confianza y decidir qué hacer con los valores perdidos, este cuadro de diálogo es el mismo para todas las pruebas T (e incluso para las pruebas no paramétricas):

- Excluir casos según análisis: Se encargará de excluir, de cada prueba que se haga, los casos que sean un valor perdido.
- Excluir casos según lista: Excluirá de las pruebas los casos con algún valor perdido en cualquiera de las variables colocadas en la casilla “Variables a contrastar”.

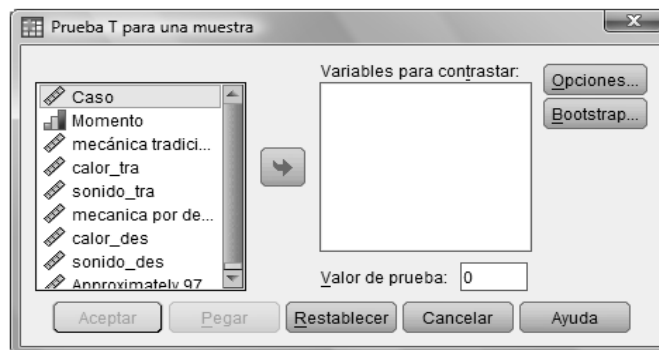


Fig. 30. Prueba T para una muestra

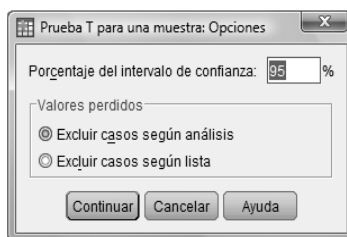


Fig. 31. Opciones

¿Dudas hasta el momento? Sigamos, como ya dijimos decidiremos aceptar o no la hipótesis nula de acuerdo al valor del estadístico de prueba y la regla de decisión, para ello buscaremos dicho valor de “T” en una tabla de la distribución t-Student, pero... la verdad no es nada “In” hacerlo.

Y... ahora, ¿quién podrá ayudarnos a decidir? SPSS arroja como resultado dos tablas, una contiene los estadísticos descriptivos y la otra los datos relativos a la prueba de hipótesis. Con la segunda tabla podremos decidir si aceptamos la hipótesis nula o no usando alguna de las anotaciones siguientes:

1. De acuerdo al valor de “Sig. (bilateral)” (nivel crítico bilateral). Éste valor se refiere al grado de compatibilidad entre valor propuesto y la información de la muestra, de modo que si éste valor es menor que 0.05 se concluye que los datos son incompatibles con la hipótesis de que el valor de la media poblacional es el propuesto.
2. Otra camino para determinarlo es mediante el intervalo de confianza, si este incluye al cero entonces diremos que no podemos rechazar la hipótesis nula.

Como ya vimos tenemos dos formas que nos ayudarán a determinar si se acepta la hipótesis nula de la prueba T. En caso de ser aceptada la forma en que debemos reportarla es: No se puede rechazar la hipótesis nula de que los datos fueron obtenidos de una población con media X.

Ambas reglas podrán ser usadas para cualquier tipo de prueba de hipótesis, con la excepción de que la conclusión cambiará de acuerdo con el tipo de prueba.

3.1.3 PRUEBA T PARA MUESTRAS INDEPENDIENTES

Con la prueba T para muestras independientes contrastamos la hipótesis de diferencia de medias. Por ejemplo, el salario promedio entre hombres y mujeres, el promedio de calificaciones de dos turnos en una escuela o si el promedio de vida es el mismo en hombres que en mujeres.

Tenemos dos poblaciones con distribución normal de las cuales construimos dos muestras aleatorias de tamaño n_1 y n_2 , mediante estas muestras obtenemos sus respectivas medias para contrastar la hipótesis de que las medias poblacionales son iguales, es decir:

$$H_0: \mu_1 - \mu_2 = \delta_0,$$

$$H_1: \mu_1 - \mu_2 \neq \delta_0$$

Al igual que en la prueba T, si se cumplen los criterios de una distribución normal, varianzas iguales pero desconocidas, entonces podemos usar el estadístico de prueba:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Sin embargo, para el caso en el que las varianzas sean iguales pero desconocidas definamos las variables aleatorias independientes, con distribuciones chi-cuadrado con n_1-1 y n_2-1 grados de libertad, siguientes:

$$\frac{(n_1-1)S_1^2}{\sigma^2} \quad y \quad \frac{(n_2-1)S_2^2}{\sigma^2}$$

Tomando en cuenta la definición de una variable aleatoria con distribución T, tenemos que el estadístico de prueba T queda de la forma siguiente con (n_1+n_2-2) grados de libertad:

$$T = \frac{\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\sigma^2}}}$$

$$\sqrt{\frac{n_1-1+n_2-1}{n_1-1+n_2-1}}$$

Aplicando la raíz y reduciendo el denominador tenemos:

$$T = \frac{\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}}$$

$$\sigma$$

Por último, aplicando la división, T nos queda de la forma siguiente:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

La regla de decisión que ocuparemos en este caso es rechazar H_0 cuando $T \leq t_{\alpha/2, n_1+n_2-2}$ o cuando $T \geq t_{1-\alpha/2, n_1+n_2-2}$.

Continuemos con la configuración de esta prueba en SPSS. En la Fig. 32 tenemos los elementos de la prueba T para muestras independientes, describamos ahora cuáles son los elementos que corresponden a cada casilla:

- Variables para contrastar: Contiene los datos sobre los cuales se desea comparar los grupos; por cada variable colocada en esta casilla se hará una prueba.
- Variable de agrupación: Dicha variable definirá los grupos que se desean comparar, el formato puede ser numérico o de cadena corta.

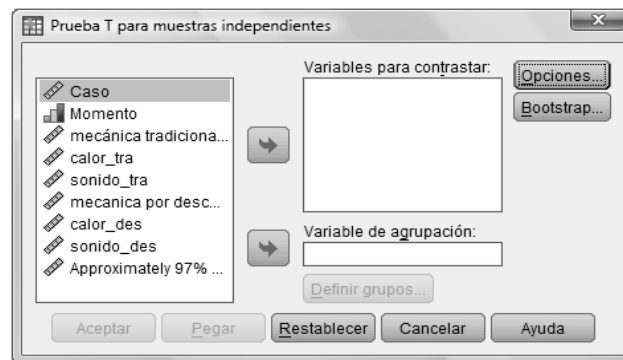


Fig. 32. Prueba T para muestras independientes

Después de haber seleccionado la variable de agrupación, se activará el botón “Definir grupos”, para definir los grupos tenemos dos opciones:

1. Usar valores especificados. El valor para cada Grupo corresponderá a los valores que ingresemos en las casillas, de tal forma que si existen otros valores no serán tomados en cuenta.

Otro aspecto importante es que si la variable es de tipo numérica los valores deben de ser enteros; sin embargo, si ésta es de tipo cadena corta escribiremos la cadena correspondiente.

2. Punto de corte. Esta opción es la ideal cuando la variable de agrupación es cuantitativa continua, de tal forma que los casos valor mayor o igual que el punto de corte serán un grupo, y el resto conformará otro grupo.

La forma en que determinaremos aceptar o no la hipótesis nula es un tanto más complicada, ya que debemos tomar en cuenta si existe igualdad de varianzas. SPSS utiliza la prueba de Levene (F) para verificarlo, la cual consiste en la comparación de las varianzas, el estadístico está definido como una variable que sigue una distribución F, la cual es la siguiente:

$$F = \frac{X/v_1}{Y/v_2}$$

Tanto X como Y son variables aleatorias independientes con distribución chi-cuadrado con n_1-1 y n_2-1 grados de libertad, respectivamente, por lo que sustituyendo sus valores en F, tenemos:

$$F = \frac{\left[\frac{(n_1 - 1)S_1^2 / \sigma_1^2}{(n_1 - 1)} \right] / (n_1 - 1)}{\left[\frac{(n_2 - 1)S_2^2 / \sigma_2^2}{(n_2 - 1)} \right] / (n_2 - 1)} = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2}$$

Sin embargo, cuando conocemos o suponemos que el valor de las varianzas es conocido, además que éstas son iguales, entonces F queda de la forma siguiente:

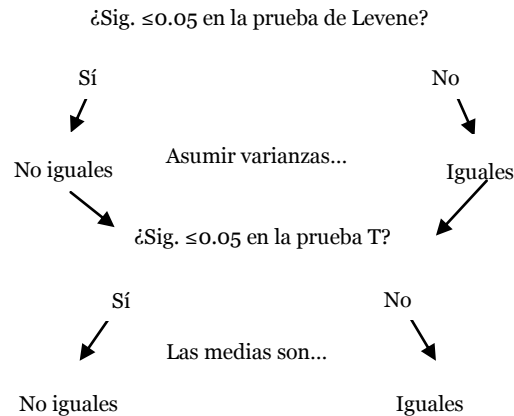
$$F = \frac{S_1^2}{S_2^2}$$

Finalmente la prueba de hipótesis es la que tenemos debajo, y rechazaremos H_0 cuando $F > f_{1-\alpha/2, n_1-1, n_2-1}$ o cuando $F \leq 1/f_{1-\alpha/2, n_1-1, n_2-1}$.

$$H_0 : \sigma_1^2 = \sigma_2^2,$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Tomando en cuenta lo anterior, utilicemos como el diagrama debajo, el cual nos permitirá decidir acerca de la prueba de Levene y de la prueba T para dos muestras independientes:



3.1.4 PRUEBA T PARA MUESTRAS RELACIONADAS

La prueba T para muestras relacionadas compara dos medias relacionadas a partir del cálculo de la diferencia entre los valores de dos variables de cada caso, de tal forma que contrasta si la media difiere de cero.

Un ejemplo muy sencillo para usar este tipo de contraste es cuando queremos probar la efectividad de un tratamiento. El procedimiento consiste en que tanto al principio como al final del mismo debemos tomar las mediciones pertinentes, hecho esto realizaremos la prueba.

Muy bien, con las muestras relacionadas (todas ellas de tamaño n) se obtiene una muestra de diferencias restando las puntuaciones de cada par, que son las n-diferencias, quedando el estadístico de prueba, con n-1 grados de libertad, definido como:

$$T = \frac{\bar{D} - \delta_0}{S_D / \sqrt{n}}$$

En donde:

$$S_D^2 = \frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n-1} \text{ y } \bar{D} = \sum_{i=1}^n \frac{D_i}{n}$$

La prueba de hipótesis queda de la forma siguiente:

$$H_0: \mu_D = \delta_0,$$

$$H_1: \mu_D \neq \delta_0$$

Rechazaremos H_0 cuando $T \leq t_{\alpha/2, n-1}$ o cuando $T \geq t_{1-\alpha/2, n-1}$.

Continuemos con la obtención de esta prueba en SPSS, una vez que estemos en el cuadro de diálogo de la Fig. 33 seleccionaremos la pareja de variables a comparar, por cada pareja que coloquemos SPSS hará una prueba.

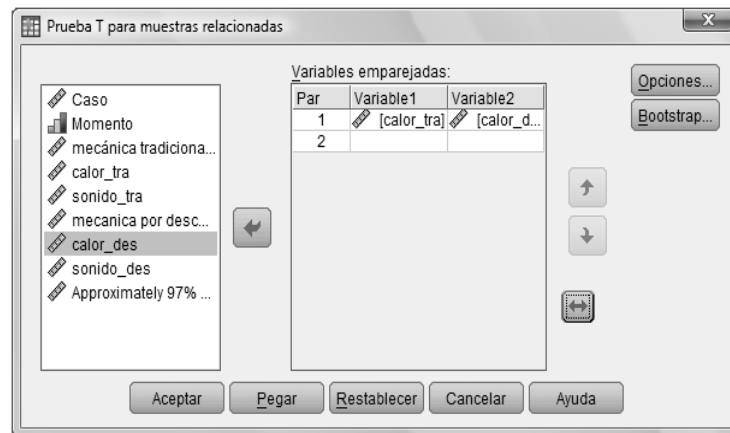


Fig. 33. Prueba T para muestras relacionadas

Como resultado tendremos tres tablas:

1. La correspondiente a los estadísticos descriptivos para cada variable.
2. Otra tabla que contiene el coeficiente de correlación de Pearson, el cual nos servirá para medir qué tanta relación existe entre los pares de variables, este tipo de análisis los veremos en el capítulo siguiente.
3. Tabla de resumen de la Prueba T para muestras relacionadas.

Si aceptamos la hipótesis nula, entonces concluiremos que no se puede rechazar la hipótesis nula de que existe igualdad de medias, y por tanto no hay una diferencia estadísticamente significativa.

3.1.5 ANOVA DE UN FACTOR

El análisis de varianza de un factor (ANOVA) compara varios grupos (más de dos), es una generalización de la prueba T para dos muestras independientes, ya que permite contrastar la hipótesis de que varias medias son iguales (hipótesis nula), el factor corresponde a la variable de agrupación (ver Fig. 34), por ejemplo un programa de incentivos, niveles socioeconómicos.

Dicha prueba de hipótesis tiene la forma:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k,$$

$$H_1 : \mu_i \neq \mu_j$$

El ANOVA maneja el supuesto de que cada grupo es una muestra aleatoria independiente procedente de una población con distribución normal con tener varianzas iguales; tomando en cuenta esto, observemos la tabla 7, en ella tenemos k grupos con n observaciones cada uno.

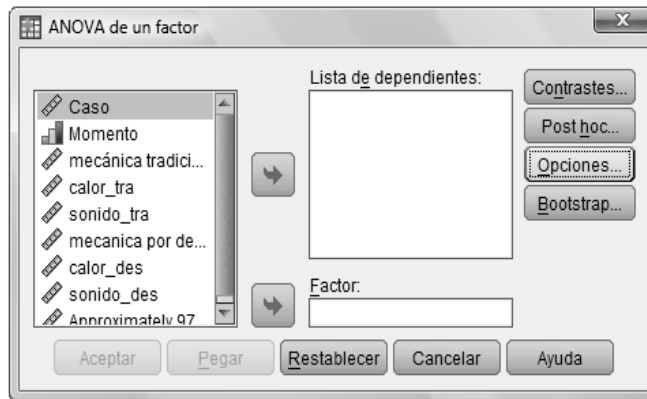


Fig. 34. ANOVA de un factor

Grupo	Observaciones	Medias muestrales	Medias poblacionales
1	$x_{11}, x_{12}, \dots, x_{1n_1}$	\bar{X}_1	μ_1
2	$x_{21}, x_{22}, \dots, x_{2n_2}$	\bar{X}_1	μ_2
...
k	$x_{k1}, x_{k2}, \dots, x_{kn_k}$	\bar{X}_1	μ_k

Tabla 7. K muestras independientes de k poblaciones

Definiremos ahora a n como la suma de todas las n_i y a las medias muestrales del grupo i y muestral global de la manera siguiente:

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} \quad \text{y} \quad \bar{x} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij}}{n} = \frac{\sum_{i=1}^k n_i \bar{x}_i}{n}$$

Muy bien, estas dos medidas nos servirán para dos tipos de variabilidades, las cuales son:

- Dentro de los grupos (intra-grupos): En el grupo i la suma de los cuadrados de las desviaciones de las observaciones respecto de su media, esto es

$$SC_i = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

La variabilidad total dentro de los grupos mide la variación en las observaciones debida a un error aleatorio, la cual está definida como:

$$SCD = SC_1 + SC_2 + \dots + SC_k = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

- Entre los grupos (inter-grupos): Mide la extensión de la variación en las observaciones, que se debe a la diferencia entre las variables (los tratamientos), la cual está definida como:

$$SCE = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2$$

Cuanto mayor es el valor de SCE mayores son las diferencias entre las medias y la media global.

Para obtener la variabilidad total que existe en los grupos solamente se suman los valores de SCD y SCE, la cual está dada de la forma siguiente:

$$SCT = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$$

Tomando en cuenta lo anterior, el estadístico de prueba determinará el grado de parentesco entre las medias comparadas, está definido por:

$$F = \frac{\frac{SCE}{(k-1)}}{\frac{SCD}{(n-k)}}$$

La regla de decisión que usaremos es rechazar H_0 cuando $F \leq f_{k-1, n-k, \alpha}$.

Gracias a esta prueba también podremos saber cuáles son las medias que difieren; para ello es necesario aplicar algún tipo de contraste ya sea a priori o a posteriori del experimento:

1. Antes (a priori). Estos contrastes permiten hacer comparaciones de tendencia y definir otro tipo de comparación entre las medias.
 - 1.1. Activando "Polinómico", se hará la comparación de la tendencia de acuerdo al origen que seleccionemos.
 - 1.2. Si personalizamos los coeficientes, estos serán contrastados mediante el estadístico t.
2. Después (a posteriori o Post hoc). Con este contraste ubicaremos en dónde se encuentra la diferencia de las medias, estas contemplan dos rubros:
 - 2.1. Asumiendo varianzas iguales: Dentro del tutorial de SPSS se recomienda elegir Tukey; aunque podemos seleccionar más de uno.
 - 2.2. No asumiendo varianzas iguales: Dentro del tutorial de SPSS se recomienda elegir Games-Howell, y también es posible seleccionar más de uno.

En las pruebas T el cuadro de diálogo de Opciones fue el mismo para todas, sin embargo, en el ANOVA de un factor, además de los Descriptivos, contamos con:

- Efectos fijos y aleatorios: Trae como resultado algunos estadísticos descriptivos tomando en cuenta efectos fijos y aleatorios.
- Prueba de homogeneidad de varianzas: Realiza la prueba de Levene para comprobar si hay homogeneidad de varianzas.
- Brown-Forsythe: Contrasta las medias por grupo para saber si existe diferencia; es el apropiado cuando no se supone igualdad de varianzas.
- Welch: Tiene la misma función que el estadístico “Brown-Forsythe”.

Con el gráfico de medias veremos las medias por subgrupos, es decir, las que tenemos definidas dentro de la variable que funge como factor.

La sección de “Valores perdidos” tiene las mismas características que en las pruebas T.

Ya sabemos para qué nos sirve cada elemento, pero, ¿en verdad tenemos que usar todo? Pues bien, esto dependerá de qué tan meticulosa requerimos que sea nuestra prueba.

3.2 PRUEBAS NO PARAMÉTRICAS

Las Pruebas T y el ANOVA, manejan el supuesto de que los datos tengan una distribución normal, además de que trabajan a partir de un parámetro (media, varianzas, etc.), es por ello que reciben el nombre de pruebas paramétricas.

Con las pruebas no paramétricas podremos hacer contrastes que no cumplan con algún requisito de las pruebas paramétricas.

SPSS cuenta con diversas pruebas no paramétricas, las encontraremos en: Analizar → Pruebas no paramétricas. Están divididas en: Una muestra, Muestras independientes, Muestras relacionadas y Cuadros de diálogo antiguos. Las tres primeras categorías están agrupadas de acuerdo al tipo de muestra, y la última tiene las pruebas por separado.

Con el fin de que nos sea más sencillo conocer estas pruebas, utilizaremos los cuadros de diálogo antiguos.

3.2.1 PRUEBA DE CHI-CUADRADO

Permite determinar si cierta distribución teórica se ajusta a la distribución de los datos, es por ello que también recibe el nombre de prueba de bondad de ajuste.

Consideremos una muestra aleatoria de tamaño n de la distribución de una variable aleatoria X dividida en K clases mutuamente excluyentes de todas las observaciones de la muestra.

La hipótesis nula es:

$$H_0 : F(x) = F_0(x)$$

$F_0(x)$ es el modelo de probabilidad propuesto de manera completa, de este modelo se puede obtener p_i , correspondiente a una observación de la i -ésima clase bajo H_0 , en donde la suma de todas las p_i es igual a 1.

Para probar esta hipótesis utilizaremos como estadístico de prueba al de Pearson, ya que compara las frecuencias observadas con las que deberíamos encontrar si dicha variable siguiera la distribución propuesta, está dado por:

$$\sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i}$$

El estadístico anterior, tiene una distribución chi-cuadrado con $k-1$ grados de libertad; el valor de N_i es la frecuencia observada de la i -ésima clase y np_i es la frecuencia que se espera bajo la hipótesis nula.

Veamos, cómo realizar esta prueba en SPSS, en la casilla de “Lista de variables a contrastar” (ver Fig. 35) por cada variable se hará un contraste. Con el botón “Exacta...” determinaremos la precisión de la prueba, así como el número de casos que se tomarán en cuenta.

En la sección de “Rango esperado”, elegiremos qué rango de valores se tomarán en cuenta, podemos elegir:

- Obtener de los datos: Considera a cada valor distinto una categoría para el análisis.
- Usar un rango especificado: Tomará en cuenta únicamente los valores Lower (Inferior) y Upper (Superior).

Dentro de la sección de “Valores esperados” especificaremos cuáles son los valores con los que deseamos comparar los valores observados.

- Todas las categorías iguales. Las frecuencias esperadas se obtienen dividiendo el número total de casos válidos entre el número de categorías de la variable.
- Valores. Son definidos por nosotros, y son interpretados como proporciones.

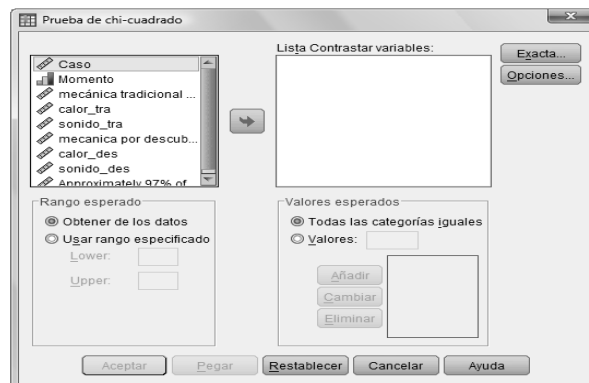


Fig. 35. Prueba de chi-cuadrado

El cuadro de diálogo de “Opciones” es el mismo para las pruebas paramétricas que para las no paramétricas.

Como resultado tendremos tres tablas, una con los estadísticos descriptivos, otra con las frecuencias y una última correspondiente a la prueba chi-cuadrado, la cual nos ayudará a decidir sobre la hipótesis nula. Si aceptamos la hipótesis nula diremos que no se puede rechazar la hipótesis nula de que los datos siguen la distribución propuesta.

3.2.2 PRUEBA BINOMIAL

La prueba binomial verifica si los datos siguen una distribución binomial, podemos relacionarla con experimentos que toman únicamente dos valores (también conocidos como valores dicotómicos), en donde uno de ellos podemos relacionarlo a un éxito y el otro a un fracaso; por ejemplo: tratado o no tratado, a favor o en contra, acierto o error.

Si extraemos muestras aleatorias de tamaño n , en la que definimos a la variable X como el número de éxitos de las n extracciones, también, si la proporción de aciertos permanece constante en cada extracción, podemos utilizar la distribución binomial para conocer la distribución exacta asociada a cada uno de los valores de X , además a medida que n aumenta la distribución de X se aproxima a una distribución normal, por lo que nuestro estadístico de prueba queda:

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

SPSS, utiliza para muestras menores a 25 la distribución binomial para obtener las probabilidades asociadas a los valores de X ; mientras que para valores mayores a 25 utiliza la distribución normal.

La “Proporción de prueba” (ver Fig. 36) se refiere a la probabilidad de que ocurra un éxito, por default es de 0.50, sin embargo, podemos adaptarlo, por tanto la probabilidad correspondiente a un fracaso será el resultado de la diferencia 1-“Proporción de prueba”.

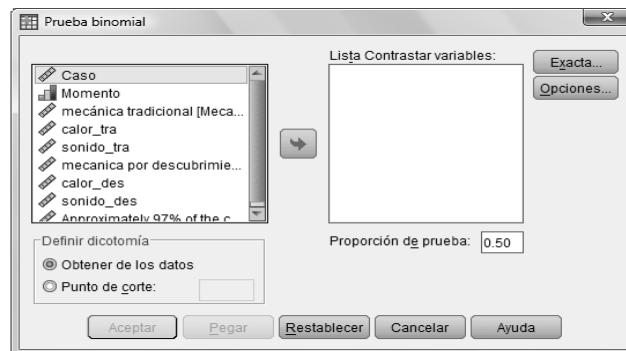


Fig. 36. Prueba binomial

Para definir los valores dicotómicos tenemos las opciones siguientes:

- Obtener de los datos. Recomendable si la variable es dicotómica.
- Punto de corte. Tiene la misma lógica de la prueba T para muestras independientes.

La regla de decisión que usaremos para esta prueba es con el valor de Sig. o con el intervalo de confianza; por lo que de aceptar la hipótesis nula concluiremos que la proporción poblacional sea superior o inferior al valor de la Proporción de prueba, es decir, no se puede rechazar la hipótesis nula de que los datos siguen una distribución binomial.

3.2.3 PRUEBA DE RACHAS

Esta prueba funge como polígrafo para comprobar si una muestra es aleatoria, esto lo hace mediante el análisis del número de rachas (R).

El término de racha se refiere a la secuencia del mismo valor, por ejemplo, supongamos que anotamos las caras (C) y cruces (X) que obtenemos al lanzar 10 veces una moneda dándonos por resultado:

CCCXCCXXXC. Pues bien, en este experimento se generaron 5 rachas: CCC, X, CC, XXX, C. Éste resultado a simple vista parece aleatorio, pero supongamos que en lugar de ello tuvimos el siguiente: CCCCCXXXXX; en éste caso fueron 2 rachas.

De nuevo usaremos una variable dicotómica, sin embargo, si se tratara de una variable cuantitativa el punto de corte (ver Fig. 37) puede ser alguna medida de tendencia central o podemos personalizar esta configuración (cuando la variable es categórica). Una vez definida esta variable, SPSS utiliza como estadístico para comprobar la hipótesis nula de aleatoriedad a:

$$Z = \frac{R - E(R)}{\sigma_R}$$

En este caso $E(R) = (2n_1n_2)/(n+1)$ y $\sigma_R = \sqrt{[2n_1n_2(2n_1n_2 - n)]/[n^2(n-1)]}$.

No olvidemos que para decidir si aceptamos la hipótesis nula usaremos el valor de Sig. o el intervalo de confianza. De aceptar la hipótesis nula, concluiremos que no se puede rechazar la hipótesis nula de que los datos sí forman parte de una muestra aleatoria.



Fig. 37. Prueba de rachas

3.2.4 PRUEBA DE K-S PARA 1 MUESTRA

La prueba K-S de una muestra (Kolmogorov-Smirnov, ver Fig. 38), es parecida a la prueba chi-cuadrado, esto es debido a que compara si la distribución de una muestra sigue una distribución Normal, Uniforme, Poisson o Exponencial.

A diferencia de la prueba chi-cuadrado, la prueba de Kolmogorov-Smirnov no requiere que los datos se encuentren agrupados, es aplicable a muestras de tamaño pequeño; ya que se basa en una comparación de las funciones de distribución acumulativa que se observan en la muestra ordenada y la distribución propuesta bajo la hipótesis nula. Si la comparación revela una diferencia suficientemente grande entre las funciones de distribución entonces la hipótesis nula se rechaza.

Nos podremos preguntar ahora, y para qué nos sirve saber qué distribución siguen nuestros datos, pues bien, gracias a esto nos será posible elaborar simulaciones, es decir, estadísticamente hablando, sabremos cómo es el comportamiento de un experimento si se realiza de nuevo pero con alguna modificación.

Continuemos, definamos a $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ como las observaciones ordenadas de una muestra ordenada de tamaño n , la función de distribución acumulativa es:

$$S_n(x) = \begin{cases} 0 & x < x_{(1)} \\ k/n & x_{(k)} \leq x < x_{(k)+1} \\ 1 & x \geq x_n \end{cases}$$

La estadística de Kolmogorov-Smirnov para esta prueba está determinado por: $D_n = \max_x |S_n(x) - F_0(x)|$, el cual tiene una distribución que es independiente del modelo propuesto en la hipótesis nula; gracias a esto, éste estadístico lo podemos evaluar en función del tamaño de la muestra y después usarlo para cualquier $F_0(x)$.

SPSS por default nos marca el contraste con una distribución Normal, sin embargo, podemos elegir alguna otra u otras. La forma para decidir es la misma que en la prueba de rachas, por lo que si se acepta la hipótesis nula diremos que no se puede rechazar la hipótesis de que los datos siguen la distribución en cuestión.



Fig. 38. Prueba K-S para una muestra

3.2.5 PRUEBAS PARA 2 MUESTRAS INDEPENDIENTES

Este procedimiento incluye cuatro pruebas, las cuales describiremos a continuación:

- U de Mann-Whitney: Es la prueba genérica de la prueba T sobre diferencia de medias en caso de que no se cumplan los supuestos de normalidad e igualdad de varianzas, o que el nivel de medida de los datos sea ordinal.

Consideremos dos muestras independientes de tamaños n_1 y n_2 extraídas de la misma población o de poblaciones idénticas. De este modo tendremos $n = n_1 + n_2$ observaciones, ahora asignaremos R_i rangos a las n puntuaciones. Ahora, definiremos a los estadísticos $S_1 =$ "suma de los rangos asignados a la muestra 1", y $S_2 =$ "suma de los rangos asignados a la muestra 2", además de considerar el estadístico U el cual para cada grupo será:

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - S_1 \text{ y } U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - S_2$$

Debido al supuesto de que ambas muestras fueron obtenidas de poblaciones idénticas, entonces podríamos esperar que U_1 y U_2 sean aproximadamente iguales, excepto en las fluctuaciones propias del azar. Si U_1 y U_2 son muy distintos, entonces existirá cierta evidencia de que las muestras proceden de poblaciones distintas.

Con lo anterior, podríamos rechazar la hipótesis nula de igualdad de medias cuando U_1 (o U_2) es demasiado grande o demasiado pequeño, para ello tomaremos en cuenta la probabilidad asociada al estadístico U :

$$\begin{aligned} U &= U_1 & \text{si } U_1 < n_1 n_2 / 2 \\ U &= U_2 & \text{si } U_1 > n_1 n_2 / 2 \end{aligned}$$

Con muestras de tamaño menor o igual que 30, SPSS ofrece el nivel crítico exacto asociado al estadístico U ; sin embargo, para muestras mayores que 30 SPSS utiliza:

$$z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2}{n(n-1)} \left(\frac{n^3 - n}{12} - \sum_i \frac{t_i^3 - t_i}{12} \right)}}$$

El término k , se refiere al número de rangos distintos en los que empates que los existen, y t_i al número de puntuaciones empatadas en el rango i .

- K-S para 2 muestras (Z de K-S): Comprueba si dos muestras proceden de la misma población a partir de la comparación de las funciones de distribución acumulada. A diferencia de la U de Mann Whitney, K-S es sensible a cualquier tipo de diferencia ya sea en medidas de tendencia central, simetría, variabilidad, etc.

Para la obtención de las funciones de distribución se asignan rangos a los valores de X_i de cada muestra. Después de lo anterior, se procede a construir la función de distribución empírica para cada valor de X_i usando: $F_j(x_i) = i/n_j$ (i se refiere al rango). A continuación se obtienen las diferencias $D_i = F_1(X_i) - F_2(X_i)$, las cuales nos servirán para nuestro estadístico de prueba, el cual es:

$$Z_{K-S} = \max_i |D_i| \sqrt{(n_1 n_2) / (n_1 + n_2)}$$

Dicho estadístico sigue una distribución normal.

- Reacciones extremas de Mose: Comprueba si hay diferencia en la dispersión o variabilidad de las distribuciones. A diferencia de la prueba de Levene, éste método puede usarse con variables ordinales.

Diremos que ha ocurrido una reacción extrema cuando en las muestras observemos un incremento muy marcado en las observaciones y en otras que sea muy poco; por ejemplo, en un programa de salud nutricional después de haber aplicado una dieta tanto en hombres como en mujeres se observa que hubo excelentes resultados para algunos hombres y apenas bueno para las mujeres.

Consideremos dos muestras aleatorias de la misma población o de dos poblaciones idénticas. Comencemos ordenando ascendentemente las $n = n_c + n_e$ (c =control y e =experimental) observaciones como si se tratara de una sola muestra, además de asignarles rangos de 1 a n , empezando con el valor más pequeño, en caso de existir empates a éstos se les asigna el rango medio.

A continuación, se calcula la amplitud⁶ del grupo control (A_c), el resultado se debe redondear al entero más próximo, sin embargo, debido a la inestabilidad del rango, Moses sugiere usar la amplitud recortada (A_r), para ello se restan los rangos correspondientes al valor más grande y al más pequeño del grupo de control, después de eliminar los r valores grandes y los r más pequeños de ese grupo sumaremos 1 y redondearemos al entero más próximo.

Tomando en cuenta lo anterior, A_r no puede ser menor que $n_c - 2r$ (ni mayor que $n - 2r$). Asimismo, si en el grupo experimental se han producido reacciones extremas, la amplitud del grupo de control tenderá a su valor mínimo. Con base en lo anterior, es interesante conocer cuáles son los valores de A_r que superen el valor de $n_c - 2r$, para ello usaremos (s es la cantidad en que un valor A_r supera a $n_c - 2r$):

$$P(A_s \leq n_c - 2r + s) = \frac{\sum_{i=0}^s \left[\binom{i + n_c - 2r - 2}{i} \binom{n_e + 2r + 1 - i}{n_e - i} \right]}{\binom{n}{n_e}}$$

SPSS calcula dicha probabilidad tanto para $r=0$ como para 0.005 ; de modo que si resulta muy pequeña rechazaremos la hipótesis de que ambas muestras proceden de poblaciones con la misma amplitud.

- Rachas de Wald-Wolfowitz. Ésta prueba es muy parecida a la prueba K-S para 2 muestras y a la de rachas. El proceso consiste en tomar ambas muestras como una sola, esto es para ordenar las observaciones y obtener el número de rachas, si existen empates SPSS calcula el número mínimo y máximo de rachas.

Nuestra hipótesis estará fundamentada en el número de rachas, si éstas son muy altas, entonces las muestras son de la misma población, caso contrario si son muy pocas rachas entonces no provienen de la misma población.

Ahora nos preguntaremos, ¿cómo determinar cuando son muy pocas o muy altas? Pues bien, si el tamaño de la muestra es mayor que 30, SPSS utiliza una aproximación Normal (ver prueba de rachas) con un nivel crítico unilateral, es decir, la probabilidad de obtener un número de rachas igual o menor que el obtenido. Cuando el tamaño sea menor o igual a 30, entonces manejaremos un nivel crítico unilateral exacto, para lo cual debemos tomar en cuenta si el número de rachas es par, entonces tenemos:

$$P(r \leq R) = \frac{2}{\binom{n}{n_1}} \sum_{r=2}^R \left[\binom{n_1 - 1}{r/2 - 1} \binom{n_2 - 1}{r/2 - 1} \right]$$

Para un número de rachas impar ($K=2r-1$):

$$P(r \leq R) = \frac{1}{\binom{n}{n_1}} \sum_{r=2}^R \left[\binom{n_1 - 1}{k-1} \binom{n_2 - 1}{k-2} + \binom{n_1 - 1}{k-2} \binom{n_2 - 1}{k-1} \right]$$

⁶ (Diferencia del máximo y mínimo)+1.

Rechazaremos la hipótesis nula de que las muestras provienen de la misma población cuando la probabilidad obtenida sea menor que 0.05.

Al igual que en la prueba de K-S para una muestra, el cuadro de diálogo para 2 muestras independientes (ver Fig. 39) nos permitirá elegir más de una prueba, para las cuales colocaremos las variables a contrastar en la casilla “Lista de variables a contrastar”, eso sí, no debemos olvidar establecer una variable de agrupación, la cual tiene la misma idea que en la prueba T para muestras independientes.

Como resultado tendremos por cada prueba: una tabla propia de ella de la prueba, y una de estadísticos de contraste la cual contiene el ya famoso, para nosotros, valor de Sig.



Fig. 39. Pruebas para 2 muestras independientes

3.2.6 PRUEBAS PARA K MUESTRAS INDEPENDIENTES

A diferencia del grupo de pruebas anteriores, las pruebas K para muestras independientes manejan más de dos grupos o muestras, y éstas son:

- H de Kruskal-Wallis: Es la extensión de la prueba Mann-Whitney, solamente que no requiere de los supuestos de normalidad y varianzas iguales, además, permite trabajar con variables ordinales.

Consideremos k muestras aleatorias independientes de tamaños n_1, n_2, \dots, n_k extraídas de la misma población o de k poblaciones idénticas. Asimismo, definamos a n como la suma de todos los tamaños de las k muestras. Para continuar, seguiremos con la asignación de los rangos de 1 hasta n como si tratara de una muestra, y en caso de empates usaremos el promedio de los rangos empatados.

Nombremos a R_{ik} como los rangos de las observaciones i de la muestra k , a R_k como la suma de los rangos asignados a las n_k observaciones de la muestra k , es decir:

$$R_k = \sum_{i=1}^{n_k} R_{ik} \text{ y } \bar{R}_k = \frac{R_k}{n_k}$$

Cuando los valores de R_k , de las distintas muestras sean parecidos entonces se cumple la hipótesis nula de que las K poblaciones son idénticas; para lo cual emplearemos el estadístico H , dado por:

$$H = \frac{12}{n(n+1)} \sum_{k=1}^K \frac{R_k^2}{n_k} - 3(n+1)$$

- De la mediana. Ésta prueba es muy parecida a la de chi-cuadrado, con la diferencia de que se dicotomiza una variable cuantitativa a partir de la mediana; tiene como hipótesis nula que las muestras provienen de poblaciones con la misma mediana.

El procedimiento consiste en ordenar todas las observaciones y calcular la mediana total bajo los criterios siguientes (X_n corresponde al valor más grande):

$$\text{Mediana} = (X_{[n/2]} + X_{[n/2+1]})/2 \text{ si } n \text{ es par}$$

$$\text{Mediana} = X_{[(n+1)/2]} \text{ si } n \text{ es impar}$$

Hecho lo anterior, obtendremos dos grupos uno en el que los valores son iguales o menores que la mediana y otro en el que son mayores. El siguiente paso que hace SPSS es obtener una tabla de contingencia de $2 \times K$ (K es el número de muestras independientes). Finalmente se aplica el estadístico Chi-cuadrado disponible en el cuadro de diálogo de Tablas de contingencia.

- Jonckheere-Terpstra. Dicha prueba es la adecuada cuando los datos tienen un orden natural, ya sea ascendente o descendente. Es importante mencionar que únicamente podemos realizar la prueba si instalamos el módulo adicional “Pruebas exactas”.

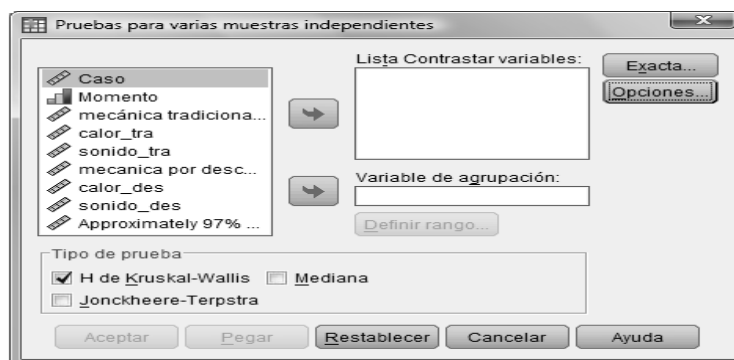


Fig. 40. Pruebas para varias muestras independientes

Después de haber elegido tanto a las variables a contrastar así como a la variable de agrupación (ver Fig. 40) será necesario que definamos el rango, todos los valores que se encuentren dentro de él se tomarán para los grupos.

Una de las tablas que arrojará como resultado tendrá correspondencia a los de la prueba y otra más a los estadísticos de contraste, en donde ubicaremos al valor de Sig.

3.2.7 PRUEBAS PARA 2 MUESTRAS RELACIONADAS

El propósito de estas pruebas es analizar si existe igualdad de medianas (Wilcoxon y Signos), igualdad de proporciones (McNemar), y los cambios de una respuesta dentro de una variable categórica (prueba de homogeneidad marginal).

La prueba de Wilcoxon consiste en tomar dos medidas a un grupo de m sujetos y calcular las diferencias en valor absoluto entre dos puntuaciones de cada par:

$$D_i = |X_i - Y_i| \quad (i = 1, 2, \dots, m)$$

Únicamente tomaremos en cuenta las diferencias nulas, a las que asignaremos R_i rangos; a partir de ellos obtendremos la suma (S_+) de las diferencias en las que $X_i > Y_i$; y también la suma (S_-) en la que las diferencias tienen $X_i < Y_i$.

Si suponemos que X_i y Y_i provienen de poblaciones con la misma mediana (hipótesis nula) debemos asumir que:

$$P(X_i < Y_i) = P(X_i > Y_i)$$

La hipótesis nula será verdadera cuando la distribución de las diferencias es simétrica, es decir, que las D_i positivas se alejarán de cero en igual medida que las D_i negativas, esto es, que:

$$S_+ = \sum R_i^* = S_- = \sum R_i^-$$

Con muestras de tamaño pequeño es fácil identificar si las diferencias son simétricas, sin embargo, para casos de muestras de tamaño grande podremos usar el estadístico Z , con distribución normal, siguiente:

$$Z = \frac{S - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \sum_{i=1}^k \frac{t_i^3 - t_i}{48}}}$$

Continuemos con la prueba de los signos. Al igual que en la prueba de Wilcoxon, manejaremos las diferencias y la probabilidad de que X_i y Y_i , provienen de poblaciones con la misma mediana, con la salvedad de que debemos verificar que sea igual a 0.5.

Bajo esta condición, así como la de la hipótesis nula, la cual es la misma que en la de Wilcoxon, las variables: n_+ =número de signos positivos, y n_- =número de signos negativos, se distribuyen binomialmente con n parámetros y $p=0.5$.

De acuerdo a lo anterior, lo siguiente es tomar en cuenta el tamaño de la muestra, si éste es menor igual a 25, SPSS toma el valor $k = \min(n_+, n_-)$ y utiliza las probabilidades de la distribución binomial. En caso contrario, si el tamaño de la muestra es mayor que 25 SPSS utiliza el estadístico:

$$Z = \frac{k + 0.5 - \frac{n}{2}}{0.5\sqrt{n}}$$

Con la prueba de McNemar contrastaremos la igualdad de proporciones antes-después, esto es, que la proporción de éxitos es la misma en la medida antes y en la medida después; por ejemplo, las intenciones de voto al inicio de las campañas y al final de las campañas; el favorito para ganar una copa de los partidos de fútbol al inicio de la temporada y en los octavos de final.

Lo anterior se logra con la comparación de los cambios que producen el antes y el después en ambas direcciones. Podremos rechazar la hipótesis de igualdad de proporciones cuando los cambios en una dirección sean significativamente más numerosos que en la otra.

Cuando se trate de cambios no muy grandes, SPSS calcula la probabilidad de encontrar un número de cambios como el observado o más alejado del valor esperado. Esto lo hace basándose en la distribución binomial con $p = 0.5$.

Si el número de cambios es muy grande, SPSS ofrece una probabilidad basada en el estadístico de McNemar y en la distribución chi-cuadrado:

$$X_{McNemar}^2 = \frac{[(\text{no. de cambios en una dirección}) - (\text{no. de cambios en la otra dirección})]^2}{\text{no. total de cambios}}$$

El estadístico anterior sigue una distribución chi-cuadrado con 1 grado de libertad.

El cuadro de diálogo correspondiente es el de la Fig. 41. Dentro de contrastar pares colocaremos las parejas que queremos contrastar, cabe señalar que tiene la misma lógica que el de la prueba T para muestras relacionadas.

Como resultado tendremos por cada prueba (excepto para la de los signos, ya que son 3 las tablas que se generan) dos tablas: una correspondiente a la prueba y otra para los estadísticos de contraste. Para decidir si aceptamos o no la hipótesis nula usaremos el valor "Sig." correspondiente.



Fig. 41. Pruebas para dos muestras relacionadas

3.2.8 PRUEBAS PARA K MUESTRAS RELACIONADAS

Este conjunto de pruebas es la generalización de las pruebas para 2 muestras relacionadas, ya que no nos limita el número de grupos o muestras.

Las pruebas que maneja SPSS en este rubro son las siguientes:

- Friedman. El escenario a analizar es el mismo que en el ANOVA, solamente que a diferencia de la prueba antes mencionada no exige los supuestos de normalidad e igualdad de varianzas, y permite trabajar con variables ordinales.

El procedimiento consiste en asignar rangos a los n bloques de la k muestras, siendo que $k(k+1)/2$ es la suma de los rangos en cada bloque.

Adicionalmente, tendremos a R_{ik} como al rango asignado al sujeto o bloque k en el tratamiento o muestra k , y a R_k como la suma de los rangos asignados a las n observaciones de la muestra k ; ambos los utilizaremos en las expresiones siguientes:

$$R_k = \sum_i^n R_{ik} \text{ y } \bar{R}_k = \frac{R_k}{n}$$

Si los promedios poblacionales son iguales, los R_k serán parecidos, de acuerdo a esto, el estadístico de Friedman se distribuye como una chi-cuadrada con $k-1$ grados de libertad:

$$X_r^2 = \frac{12}{nk(k+1)} \sum_k R_k^2 - 3n(k+1)$$

- W de Kendall. Es la normalización de la prueba de Friedman, para la cual estableceremos a

$$R_i = \sum_{k=1}^k R_{ik}$$

como a la suma de los rangos correspondientes al sujeto u objeto i

Diremos que se da concordancia perfecta entre los k conjuntos de rangos cuando se valoran igual a los n sujetos u objetos, o cuando los n sujetos u objetos son clasificados de manera idéntica en las k características. Lo anterior significa que coincide la asignación del rango 1 a uno los sujetos u objetos, el rango 2 a otro sujetos u objetos, ..., el rango n a otro de los sujetos. El efecto que esto tiene en los totales R_i es que los totales correspondientes a los diferentes sujetos u objetos serán: $1J, 2J, 3J, \dots, iJ, \dots, nJ$.

Por el contrario, cuando no existe concordancia entre los k conjuntos de rangos, los n sujetos u objetos son valorados de forma distinta, quedando los totales R_i como:

$$R_1 = R_2 = \dots = R_i = \dots = R_n = \frac{k(n+1)}{2}$$

Así pues, el grado de concordancia queda reflejado en la variabilidad entre los totales R_i de los diferentes sujetos u objetos.

Ahora bien, cuando la concordancia entre los k conjuntos es perfecta, la variabilidad entre los R_i es máxima; cuando la concordancia es nula, la variabilidad entre los R_i es mínima. Tomando en cuenta lo anterior, definiremos al estadístico S como:

$$S = \sum_{i=1}^n \left(R_i - \frac{k(n+1)}{2} \right)^2$$

S mide la variabilidad observada contra la total R_i y el total que cabría esperar si la concordancia fuera nula, alcanzará su valor máximo en el caso de concordancia perfecta, es decir, cuando entre los R_i totales exista la máxima variabilidad dada por la expresión:

$$S_{\text{máx}} = \frac{k^2 n (n^2 - 1)}{12}$$

Si queremos obtener un coeficiente que valga cero en el caso de concordancia nula y 1 en el caso de concordancia perfecta, transformaremos a S dividiéndolo entre su valor máximo posible, de la cual obtendremos el coeficiente de concordancia de Kendall:

$$\hat{W} = \frac{12 \sum_i R_i^2}{k^2 n (n^2 - 1)} - \frac{3(n+1)}{n-1}$$

Con el coeficiente anterior, nos ayudaremos para la construcción de un estadístico de prueba adecuado para hacer inferencias de éste mismo, ya que es posible transformarlo en el estadístico siguiente:

$$X_r^2 = k(n-1)\hat{W}$$

- Q de Cochran. Es la generalización del estadístico de McNemar, teniendo como variable dependiente a una dicotómica.

Las proporciones marginales P_{*j} representan las proporciones de aciertos de cada muestra, las cuales están definidas como: $P_{*j}=T_{*j}/n$, con T_{*j} como la suma de aciertos de cada muestra. Si las k muestras provienen de poblaciones idénticas, entonces las P_{*j} serán iguales, esto es a partir del estadístico de prueba:

$$Q = \frac{k(k-1)\sum T_{*k}^2 - (k-1)T^2}{kT - \sum T_{*k}^2}$$

Q se distribuye como una chi-cuadrado con $k-1$ grados de libertad.

Nuevamente, las variables que queramos estén involucradas en la (s) pruebas (s) debemos colocarlas en la casilla “Variables de contraste” (ver Fig. 42).

Las tablas resultantes serán una con los estadísticos de contraste y el resto dependerá del tipo de prueba elegida. Al igual que en el resto de la pruebas vistas, la decisión de aceptar o no la hipótesis nula dependerá del valor obtenido en “Sig.”

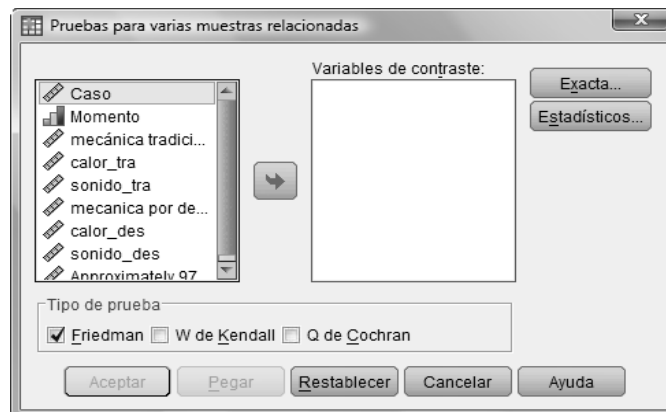


Fig. 42. Pruebas para varias muestras relacionadas

3.3 EJEMPLO GENERAL

Esta parte de la estadística inferencial tiene amplias aplicaciones, por ejemplo, en la psicología, la medicina, dentro de una empresa, etc.

En esta ocasión construiremos nuestra información. Imaginemos que nos encontramos en una preparatoria en la que se está estableciendo un nuevo sistema de enseñanza virtual en las materias de Inglés, Matemáticas y Filosofía. De modo que para probar su efectividad se dividió a los grupos en dos secciones. En una de ellas se maneja el sistema virtual y en otra la forma tradicional de impartir la clase.

Antes de continuar, de acuerdo a todas pruebas vistas a lo largo del capítulo, estableceremos los supuestos de que los datos que vamos a generar siguen una distribución normal y tienen varianzas iguales.

Nos ayudaremos de la función “aleatorio.entre (inferior, superior)” del programa Excel para construir nuestro archivo. La distribución de los valores inferior y superior de la fórmula mencionada los tenemos en la tabla a continuación mostrada:

Variable	Inferior	Superior
Edad	17	21
Género	0 (Masculino)	1 (Femenino)
Turno	1 (Matutino)	2 (Vespertino)
Ingl_vir, Ingl_aul, Mate_vir, Mate_aul, Fil_vir, Fil_aul	5	10

Tabla 8. Variables

Bastará con la primer fórmula de cada variable para que obtengamos 500 datos de cada una.

Con nuestro estudio supuesto averiguaremos qué efecto tuvieron los dos programas de estudio en los grupos vespertino y matutino, así como en mujeres y hombres, eso sí, sin olvidar cómo influye la edad de los alumnos.

Nuestra hipótesis para este ejemplo será que ambos métodos tienen el mismo resultado en todas las materias para las cuales se realizó el experimento.

	Masculino				Femenino			
	Matutino		Vespertino		Matutino		Vespertino	
	Media	Desv. típ.	Media	Desv. típ.	Media	Desv. típ.	Media	Desv. típ.
Ingl_vir	7.70	1.851	7.30	1.691	7.75	1.704	7.15	1.741
Ingl_aul	7.70	1.662	7.85	1.757	7.67	1.816	7.49	1.749
Mate_vir	7.34	1.642	7.35	1.532	7.51	1.751	7.53	1.639
Mate_aul	7.55	1.743	7.58	1.582	7.39	1.877	7.42	1.647
Filo_vir	7.79	1.701	7.59	1.741	7.40	1.757	7.50	1.546
Filo_aul	7.44	1.804	7.53	1.713	7.61	1.775	7.46	1.726

Tabla 9. Informe

Utilizaremos el procedimiento “Medias” para comparar los grupos conformados por mujeres y hombres con respecto a los turnos que maneja nuestra institución. Dentro de la casilla de “Dependientes” colocaremos todas las variables de las materias, y en “Independientes”, en una primera capa a la variable Género y en otra capa “Turno”.

La tabla anterior es el Informe, cabe mencionar que está editada para una mejor presentación. A partir de ella podemos hacer alguna de las siguientes conclusiones:

1. La calificación promedio de los hombres tanto en el sistema Inglés Virtual como el de Aula, en el turno matutino fue de 7.7.

2. El programa de Matemáticas Virtual tuvo mayor aceptación con las mujeres de ambos turnos.
3. La materia de Filosofía en el sistema tradicional tuvo mejor promedio de calificaciones en las alumnas del turno matutino.

Gracias al análisis anterior visualizamos los promedios de acuerdo a dos grupos, pero en general no sabemos cuáles fueron los resultados. Para esta parte ocuparemos la Prueba T con un valor de prueba de 7.5, la cual analizaremos a continuación.

De acuerdo con la tabla 10, en todas las materias, excepto en la de inglés en el sistema tradicional, se acepta la hipótesis nula de la prueba T con el valor de prueba de 7.5, es decir, no se puede rechazar la hipótesis de que los datos provienen de una población en la que la media es de 7.5, que es un valor aceptable por tratarse de la primera vez en la que se pone en práctica la dinámica de la enseñanza en la forma virtual.

Valor de prueba = 7.5						
	T	Gl	Sig. (bilateral)	Diferencia de medias	95% Intervalo de confianza para la diferencia	
					Inferior	Superior
Ingl_vir	-.330	499	.742	-.026	-.18	.13
Ingl_aul	2.229		.026	.174	.02	.33
Mate_vir	-.926		.355	-.068	-.21	.08
Mate_aul	-.235		.814	-.018	-.17	.13
Filo_vir	.928		.354	.070	-.08	.22
Filo_aul	.102		.919	.008	-.15	.16

Tabla 10. Prueba para una muestra

El siguiente paso es verificar cómo fue el comportamiento en los dos turnos, para ello ocuparemos la Prueba T para muestras independientes. La Tabla 11 tiene únicamente el contraste de la materia de inglés en ambos sistemas de acuerdo a los dos turnos que manejamos.

Como ya habíamos mencionado en el tema 3.1.3, será necesario que veamos primero qué pasa con la prueba de Levene. Dado que el valor de significancia bilateral, en ambos casos, es mayor que 0.05, entonces se asumen varianzas iguales.

El siguiente paso es analizar los resultados de la Prueba T. En el caso de Inglés Virtual se rechaza la hipótesis de que el promedio de calificaciones es el mismo en los dos turnos, mientras que en el caso de Inglés en el aula se acepta la hipótesis de que en promedio las calificaciones de los alumnos en ambos turnos son iguales.

Ya hemos visto que el nuevo modelo de enseñanza tuvo buena aceptación, además de saber cómo fue su impacto en los turnos disponibles en la materia de Inglés, sin embargo, aún nos falta saber cuál fue la efectividad de ambos métodos, para ello usaremos la Prueba T para muestras relacionadas con la materia de matemáticas.

De acuerdo a los resultados de la prueba (ver Tabla 12) podemos concluir que las calificaciones de la materia de matemáticas virtual son estadísticamente no significativas, de tal forma que da igual un método que otro.

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	95% Intervalo de confianza para la diferencia	
									Inf	Sup
Ingl_vir	Se han asumido varianzas iguales	1.333	.249	3.203	498	.001	.500	.156	.193	.807
	No se han asumido varianzas iguales			3.204	497.816	.001	.500	.156	.194	.807
Ingl_aul	Se han asumido varianzas iguales	.214	.644	.161	498	.872	.025	.156	-.282	.332
	No se han asumido varianzas iguales			.161	497.585	.872	.025	.156	-.282	.332

Tabla 11. Prueba T de muestras independientes

		Par 1		
		Mate_vir - Mate_aul		
Diferencias relacionadas	Media	-.050		
	Desviación típ.	2.341		
	Error típ. de la media	.105		
	95% Intervalo de confianza para la diferencia	Inf	-.256	
		Sup	.156	
T	-.477			
Gl	499			
Sig. (bilateral)	.633			

Tabla 12. Prueba T para muestras relacionadas

Cabe mencionar que no necesariamente ocupamos todas las pruebas existentes, ya que en este caso no podemos usar el ANOVA de un factor debido a que tenemos únicamente dos grupos por variables, ya sea por turno o por género o por tratamiento, si tuviéramos una tercera metodología de enseñanza sí es una buena opción usar el ANOVA.

Este ejercicio fue un esbozo de todo el análisis que se puede hacer con estas pruebas; sin embargo, dado que construimos nuestros datos, chequeemos con la prueba de rachas qué tan aleatorios son.

Tomando en cuenta la Tabla 13, podemos decir que todas las variables, exceptuando la de turno (debido a que tuvo únicamente una racha), son aleatorias, por lo que Excel si arroja valores aleatorios.

Nos apoyaremos de SPSS para crear una variable que haga referencia a la llegada de los alumnos. Dentro de crear variable elegiremos la categoría “Números aleatorios” y luego buscaremos a “RV.Poisson”, una vez seleccionada, introduciremos 5 y luego pulsaremos en Aceptar. Con esto tendremos las 500 variables correspondientes a la llegada.

	Valor de prueba	Casos < Valor de prueba	Casos >= Valor de prueba	Casos en total	Núm. de rachas	Z	Sig. asintót. (bilateral)
Edad	19	190	310		244	.703	.482
Género	1	246	254		238	-1.158	.247
Ingl_vir	7	177	323		225	-.459	.647
Turno	1	0	500		1		
Ingl_aul	8	225	275	500	251	.226	.821
Mate_vir	7	162	338		211	-.922	.356
Mate_aul	7	173	327		241	1.357	.175
Filo_vir	8	242	258		237	-1.232	.218
Filo_aul	7	169	331		211	-1.376	.169

Tabla 13. Prueba de rachas

La distribución de Poisson describe el número de eventos que ocurren a una velocidad constante, su ejemplo más común son las líneas de espera.

Ahora usemos la prueba K-S (ver Tabla 14) para una muestra para comprobar que la llegada tiene una distribución de Poisson, así como la distribución de las calificaciones de Filosofía siguen una distribución Normal.

La materia de Filosofía en sus dos presentaciones no tiene una distribución normal, de tal forma que para comparar su efectividad lo adecuado es usar la prueba para dos muestras relacionadas. Por otra parte, comprobamos que efectivamente las variables que creamos con la función RV.Poisson siguen dicha distribución.

		Filo_vir	Filo_aul	lleg_poi
N		500	500	500
Parámetros	Media	7.57	7.51	4.8820
	Desv. típica	1.687	1.752	
Diferencias más extremas	Absoluta	.156	.147	.013
	Positiva	.136	.143	.012
	Negativa	-.156	-.147	-.013
Z de K-S		3.480	3.283	.283
Sig. asintót. (bilateral)		.000	.000	1.000

Tabla 14. Prueba K-S para una muestra

La Tabla 15 es el resultado de la prueba de Wilcoxon. Dado que el valor de Significancia es mayor que 0.05, concluiremos que los promedios no difieren significativamente, es decir, que tiene el mismo efecto

usar un estilo de aprendizaje que otro, sin embargo, tomemos en cuenta que es la primer vez en la que se utiliza el modelo Virtual.

En resumen, concluiremos que para este experimento los métodos de enseñanza no resultaron exactamente iguales con respecto a las calificaciones de los alumnos, sin embargo, estadísticamente hablando, para la materia de Matemáticas no hubo diferencia significativa de tal forma que dio igual usar un método que otro; para el caso de las otras materias tendríamos que hacer la correspondiente prueba T para muestras relacionadas.

Rangos				Estadísticos de contraste	
	N	Rango promedio	Suma de rangos	Filo_aul - Filo_vir	
Rangos negativos	215	217.18	46693	Z	-0.657
Filo_aul - Filo_vir Rangos positivos	209	207.69	43407	Sig. asintót. (bilateral)	0.511
Empates	76				
Total	500				

Tabla 15. Prueba de Wilcoxon

3.4 EJERCICIO COMPLEMENTARIO

Este ejercicio consistirá en revisar un artículo de la revista chilena *Scielo*, la cual data del año 2002 y lleva por título “Epidemiología de los accidentes en la infancia de la Región Centro Cuyo”.

Dicho artículo, es un estudio hecho por varios Doctores dedicados a la pediatría, la dirección electrónica en donde podemos encontrarlo es: <http://www.scielo.org.bo/pdf/rbp/v41n2/v41n2a11.pdf>.

El objetivo de este artículo es: “Establecer la prevalencia de accidentes en niños atendidos en servicios de guardia, describir sus características, determinar asociaciones entre las variables estudiadas y la posibilidad de accidentarse”.

A simple vista pareciera no ser un tema estadístico, sin embargo, no podemos confiarnos de nuestro sentido común para emitir conclusiones que pueden estar involucradas en aspectos importantes.

La población abarcó a pacientes de 0 a 14 años atendidos durante la primera quincena de agosto y diciembre de 1998 en 17 servicios de guardia de hospitales públicos y privados.

La metodología usada para la redacción y elaboración de este artículo es la estándar, por lo que aquí no sólo tenemos los datos y ya, sino que será necesario delimitar nuestra población, identificar las variables a estudiar, establecer los objetivos (general y específicos en su caso), y plantear qué análisis son los que vamos a ocupar, así como la forma de obtener los datos.

Concentrémonos en revisar cómo fue que eligieron usar la prueba T, el ANOVA y las pruebas paramétricas:

1. Prueba de homogeneidad de los datos, corresponde a la categoría de no paramétricas para 2 muestras relacionadas. La razón para ocuparla es que los datos no cumplían los supuestos de normalidad.

2. Prueba ANOVA, la tenemos dentro de las paramétricas; en este caso el factor correspondió a una institución y estacionalidad, mientras que la variable dependiente fue el número de ocurrencia de accidentes.
3. Prueba T, no especifican cuál de ellas, pero por la forma en que la usaron debe tratarse de la Prueba T para muestras relacionadas, que es la que se encarga de encontrar diferencias significativas.
4. Kruskal-Wallis y K-S, son pruebas paramétricas que analizan dos o más muestras provenientes de la misma población, con esto establecieron si los datos de distintas instituciones siguen la misma distribución.
5. Prueba Chi-cuadrado, es no paramétrica, con la que comprobamos si una distribución teórica se ajusta a nuestros datos, en este caso más que ajustarse la usaron para establecer un nivel de asociación.

El estudio es realmente interesante, además de que maneja un sustento estadístico, pero, ¿qué conclusiones podemos anexarle? Veamos:

1. Los hospitales “medianos” mantienen una tendencia similar en la ocurrencia de accidentes, sería interesante investigar si es debido a la capacidad, tanto humana como de instalaciones, o de la zona en la que se encuentra.
2. Dentro del rubro de los niños accidentados los varones representan la mayoría con un 62.5%, con una edad promedio de 5.64.
3. A pesar de la inseguridad en las calles, sorprendentemente el lugar con mayor ocurrencia de accidentes es el hogar con 51.9%.
4. Con lo que respecta a la gravedad de los accidentes, un 72.4% fueron leves.
5. La incidencia de una caída no depende de la edad.
6. El número promedio de accidentes de tránsito no difiere significativamente en los hospitales analizados.

En muchas ocasiones un analista no solamente debe de concluir, sino también proponer líneas de acción, algunas de ellas las podemos ver dentro de la “DISCUSIÓN”, sin embargo, valdría la pena realizar un estudio relativo al servicio prestado en cada hospital, por ejemplo: con cuánto personal se cuenta en cada turno, qué se hace en caso de que no haya un medicamento, cuáles son sus medidas de seguridad, las instalaciones son de calidad. Incluso anexar una pregunta con la que se mida cuál es la hora (o turno) del día en la que llegue más gente al hospital. De este modo se puede complementar este estudio, además de usar más pruebas estadísticas que sustenten nuestros resultados y conclusiones.

3.5 CASO DE ESTUDIO (CONTINUACIÓN...)

Muy bien, en el capítulo 2 expusimos la forma en que se evaluó la práctica que se utilizó tanto para el grupo que usó SPSS así como para el que no, además se mencionó la elección de las pruebas estadísticas a obtener, por lo que ahora toca revisar cuáles fueron los resultados.

La prueba K-S para una muestra dio como resultado la tabla de la página siguiente.

De acuerdo con el valor de Sig., en efecto los datos siguen una distribución normal, de modo que, podemos usar una prueba T para muestras independientes, la cual servirá para saber si existe diferencia significativa en los promedios de las calificación y el tiempo invertidos de ambos grupos.

		Promedio calificaciones	Tiempo invertido
N		12	12
Parámetros normales ^{a,b}	Media	6.416667	62.7500
	Desviación típica	2.8561841	30.33188
Diferencias más extremas	Absoluta	.165	.196
	Positiva	.140	.176
	Negativa	-.165	-.196
Z de Kolmogorov-Smirnov		.572	.680
Sig. asintót. (bilateral)		.899	.745

a. La distribución de contraste es la Normal.

b. Se han calculado a partir de los datos.

Tabla 16. K-S de los grupos

En la tabla 17, se encuentran los resultados de la prueba T para muestras independientes, de acuerdo a la prueba de Leve (F), para ambas pruebas se debe asumir varianzas iguales.

Continuando con la prueba T para muestras independientes, en las dos pruebas se obtuvo un valor de Sig. mayor que 0.05, de modo que, no se puede rechazar la hipótesis de que los promedios de las calificaciones obtenidas y el tiempo invertido, tanto para el grupo que usó software y el que no lo usó son estadísticamente iguales.

Lo anterior quiere decir, que a pesar de que se usaron diferentes instrumentos de aprendizaje, no hubo una diferencia estadísticamente significativa, esto no necesariamente implica que no sea viable utilizar éste tipo de material y dinámica para el aprendizaje e interpretación de la estadística, ya que, de acuerdo a la prueba, en promedio ambos métodos dieron los mismos resultados.

Adicionalmente, analizamos los casos extremos de cada uno de los grupos, es decir, aquellos resultados en los que se detectó un alto o bajo promedio de calificaciones y de tiempo invertido, la prueba de hipótesis adecuada es la de reacciones extremas de Moses, los resultados de ésta se encuentran en la tabla 18.

		Promedio calificaciones		Tiempo invertido		
		Se han asumido varianzas iguales	No se han asumido varianzas iguales	Se han asumido varianzas iguales	No se han asumido varianzas iguales	
Prueba de Levene para la igualdad de varianzas	F	.151		3.301		
	Sig.	.706		.099		
Prueba T para la igualdad de medias	t	.904	.904	-.282	-.282	
	gl	10	9.980	10	6.999	
	Sig. (bilateral)	.387	.387	.783	.786	
	Diferencia de medias	1.5033333	1.5033333	-5.16667	-5.16667	
	Error típ. de la diferencia	1.6628854	1.6628854	18.29405	18.29405	
	95% Intervalo de confianza para la diferencia	Inferior	-2.2018062	-2.2028309	-45.92835	-48.42655
		Superior	5.2084728	5.2094976	35.59501	38.09322

Tabla 17. Prueba T: promedio calificaciones y tiempo invertido

Recordando, la prueba de Moses construye las amplitudes mediante la asignación de rangos; después se calcula la probabilidad de obtener una amplitud como esa o menor, tanto para el grupo control observado y recortado, siguiendo esta idea, la amplitud para el grupo de control fue de 11 con una probabilidad de obtener esa amplitud o menor de 0.773 para el promedio de calificaciones y de 0.273 para el tiempo invertido. Por lo que se refiere al valor de la amplitud recortada fue de 9 con una probabilidad de 0.879 para el promedio de calificaciones y de 0.5 para el tiempo invertido; de este modo, dado que en ambos casos fue mayor que 0.05 podemos considerar que no se produjeron reacciones extremas.

	Promedio calificaciones	Tiempo invertido
Amplitud observada del grupo control	11	9
Sig. (unilateral)	.773	.273
Amplitud recortada del grupo control	8	6
Sig. (unilateral)	.879	.500
Valores atípicos recortados de cada extremo	1	1

a. Prueba de Moses

b. Variable de agrupación: ID

Tabla 18. Prueba de Moses

Bien, hasta el momento simplemente se interpretaron los resultados de las pruebas de hipótesis, sin embargo, no se ha analizado cuál fue la opinión de los integrantes de cada grupo, ésta se revisó en el capítulo siguiente.

Capítulo 4. CORRELACIÓN Y REGRESIÓN LINEAL

Este hombre utiliza la estadística de la misma manera que un borracho utiliza una farola, más para apoyarse que para iluminarse.

Anónimo

Objetivo: En este capítulo obtendremos el tipo de correlación que hay en una relación de datos para asociarle un coeficiente, además de establecer un modelo de regresión lineal.

4.1 CORRELACIÓN

Con la correlación determinaremos el grado de relación que hay entre las variables (dos o más), de esta forma podremos concluir si una variable explica a otra. Algunos ejemplos son: si la estatura depende de la edad, o si el nivel de colesterol está relacionado con alimentos altos en grasas y por poco ejercicio.

Los tipos de correlación (ver Fig. 43) son:

1. Positiva: Los valores de las dos variables varían de forma parecida.
2. Negativa: En este tipo de correlación los valores de las variables varían justamente al revés.
3. Nula: No existe alguna relación entre las variables.

Dentro del menú Analizar→Correlaciones, tenemos los análisis correspondientes a este tema.

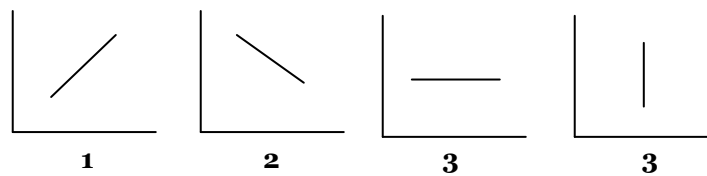


Fig. 43. Tipos de correlación

4.1.1 CORRELACIÓN BIVARIADA

Lleva este nombre debido al número de variables involucradas, en este caso se trata de dos variables. SPSS maneja tres coeficientes, los cuales son:

- Pearson. Está representado por la letra **r**, es el adecuado cuando tenemos variables cuantitativas distribuidas normalmente, la fórmula que utiliza es:

$$r_{xy} = \frac{\sum x_i y_i}{n S_x S_y}$$

En donde:

x_i, y_i : Puntuaciones diferenciales de cada grupo

n : Número de casos

S_x, S_y : Desviaciones típicas de cada variable

- Spearman: Está basado en el coeficiente de Pearson, después de transformar las puntuaciones en rangos. Es el indicado cuando los datos no siguen una distribución normal o tienen categorías ordenadas.
- Tau-b de Kendall: Es el adecuado para estudiar la relación entre variables que sean ordinales o de rangos, éste toma en cuenta el número de empates.

Los tres coeficientes toman valores entre -1 (relación lineal perfecta negativa) y 1 (relación lineal perfecta); si llegara a tomar el valor 0, esto indica que la correlación es nula.

Antes de realizar cualquier análisis de correlación debemos tomar en cuenta que puede haber datos atípicos, para identificarlos usaremos el gráfico de caja. Asimismo, como una ayuda previa, obtendremos el diagrama de dispersión de los datos y visualizaremos de forma gráfica qué tipo de correlación tienen.

Veamos ahora cuál es el procedimiento para la correlación bivariada. En la casilla de variables (ver Fig. 44) debemos depositar las variables de las que queremos saber la correlación, después indicar los coeficientes, y la prueba de significación, la cual es una prueba de hipótesis con hipótesis nula de que el valor de r es cero, su estadístico de prueba es:

$$T = \frac{r_{xy} \sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

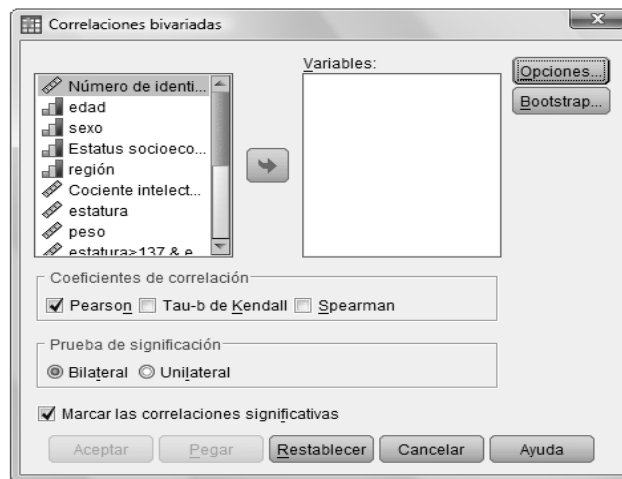


Fig. 44. Correlaciones bivariadas

Para que sepamos cómo debe ser la prueba, debemos tomar en cuenta la dirección que tiene la relación, además de lo siguiente:

- Bilateral: Indica la probabilidad de obtener coeficientes tan alejados de cero o más que el valor obtenido.

- Unilateral: Indica la probabilidad de obtener coeficientes tan grandes o más grandes que el obtenido si el coeficiente es positivo, o tan pequeño más pequeño que el obtenido si el coeficiente es negativo.

Las “Opciones” que ofrece son las siguientes:

1. Estadísticos: Es posible elegir una o ambas opciones si el coeficiente es el de Pearson.
 - 1.1. Medias y desviaciones típicas. Para cada variable se tienen estos estadísticos, además del número de casos válidos.
 - 1.2. Productos cruzados diferenciales y covarianzas. Son por cada pareja de variables.
2. Valores perdidos
 - 2.1. Excluir casos según pareja. Se excluyen los valores perdidos de una o ambas variables por cada pareja que se haga.
 - 2.2. Excluir casos según lista. De todos los análisis se excluyen todos los valores perdidos de cada variable.

Como resultado SPSS genera una tabla de estadísticos y otra de tamaño $n \times n$, en donde n es el número de variables que están involucradas en el análisis de correlación. En esta última tabla al colocar la marca de verificación en “Marcar las correlaciones significativas” en el cuadro de diálogo de la Fig. 44, SPSS coloca una marca en las parejas de variables que tengan mayor correlación. Además, en la tabla de correlaciones también se incluye el valor de “Sig.” con el que decidiremos si hay correlación diferente de cero.

4.1.2 CORRELACIÓN PARCIAL

La correlación parcial, a diferencia de la bivariada, toma en cuenta más de dos variables dentro de la relación.

El cuadro de diálogo que maneja es muy parecido al de la Fig. 45, excepto que tiene una casilla para las variables que afecten a la correlación. El número de variables que pueden estar involucradas en este tipo de correlación es de hasta 400, de las cuales, la cuarta parte de ellas pueden estar involucradas en el efecto.

Las opciones con las que cuenta el cuadro de diálogo de la correlación parcial son las mismas que en la correlación bivariada, excepto que en esta ocasión no hay coeficientes a elegir, ya que por tratarse de más de un factor el cálculo del coeficiente de correlación es más complicado.

Las tablas resultantes son las mismas que en la correlación bivariada.

4.2 REGRESIÓN LINEAL

Uno de los propósitos de establecer si existe relación entre dos o más variables es la elaboración de un modelo con el que podamos estimar a una de las variables. La que se estima mediante el modelo la llamaremos variable dependiente, mientras que la variable con la que obtenemos el modelo le llamaremos variable independiente.

La regresión lineal es una de las técnicas con las que se establece un modelo de estimación. La palabra “regresión” fue usada por primera vez, dentro del contexto de estimación, por Francis Galton [9] en sus estudios acerca de la herencia, ya que notó que las características promedio de la siguiente generación de un grupo en particular tendían a moverse en dirección de las características promedio de la población general, esta tendencia fue referida como la regresión hacia la media de la población.

Este tipo de modelos es el adecuado cuando los datos tienen una tendencia lineal, de modo que debemos estimar los parámetros que intervienen en el modelo. Existen dos tipos de regresión de acuerdo al número de variables involucradas.

Con la regresión lineal podemos establecer un modelo en el que hay una variable explicada por otra variable, mientras que en la regresión múltiple estará involucrada más de una variable para explicar a una sola variable.

El menú para la regresión lo encontraremos en: Analizar→Regresión→Lineales.

4.2.1 REGRESIÓN LINEAL SIMPLE

La regresión lineal simple es la forma más simple con la que podemos establecer una regla de correspondencia entre dos conjuntos de datos, es de gran utilidad y no requiere un análisis tan profundo.

Esta técnica consiste en ajustar una línea recta que defina una regla de correspondencia de la variable dependiente con la independiente, dada por la ecuación siguiente:

$$\hat{Y} = B_0 + B_1x$$

En donde:

B_0 : Parámetro constante que puede tomar cualquier valor, incluso cero o valores negativos, este es comúnmente llamado ordenada al origen.

B_1 : Parámetro que determina la pendiente de la regresión.

Para estimar los parámetros se utilizan las fórmulas siguientes:

$$B_0 = \frac{(\sum x)(\sum x^2) - (\sum x)(\sum xy)}{n\sum x^2 - (\sum x)^2} \text{ y } B_1 = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

En donde:

n: número de parejas de valores

$\sum x$: Suma de los valores correspondientes a la variable independiente

$\sum y$: Suma de los valores correspondientes a la variable dependiente

$\sum xy$: Suma de la multiplicación de cada valor de x por cada valor de y

$\sum x^2$: Suma de los valores correspondientes a la variable independiente elevados al cuadrado

Antes de construir el modelo de regresión lineal nos será de utilidad hacer los pasos siguientes:

- Obtener el gráfico de caja; si existen datos atípicos darles el tratamiento necesario.

- Obtener el coeficiente de correlación, para determinar si las variables involucradas en la regresión lineal son los adecuados.
- Analizar si la relación a establecer con el modelo de regresión lineal tiene lógica, ya que puede tener un coeficiente de correlación alto, pero lo que exprese no sea congruente.

En la Fig. 45 tenemos el cuadro de diálogo para la regresión lineal, si se trata del caso simple debemos colocar una variable en la casilla de “Dependientes” y una en la de “Independientes”.



Fig. 45. Regresión lineal

Con el botón “Estadísticos”, elegiremos uno o más de acuerdo a nuestras necesidades, veamos:

1. Coeficientes de regresión:

1.1. Estimadores. Guarda los coeficientes de regresión en un conjunto de datos.

1.2. Intervalos de confianza. Nos da la oportunidad de establecer el nivel deseado, aunque por default es el 95%.

1.3. Matriz de covarianzas. A diferencia de lo visto anteriormente, en cuanto a estadísticos, la covarianza involucra dos variables, se trata de una varianza conjunta.

2. Residuos:

2.1. Durbin-Watson. Es la prueba de hipótesis ideal cuando en los residuos de observa que hay autocorrelación.

2.2. Diagnósticos por caso. Es ideal para los casos que cumplan el criterio de selección.

3. Otros:

3.1. Ajuste del modelo. Presenta una lista de las variables introducidas y eliminadas del modelo y muestra los estadísticos de bondad de ajuste: R múltiple, R^2 y R^2 corregida, error típico de la estimación y tabla de análisis de la varianza.

3.2. Cambio en R cuadrado. Resume las variaciones que tiene dicho estadístico cuando se añade o elimina una variable independiente.

3.3. Descriptivos. Ofrece el número de casos válidos, media, desviación, y la matriz de correlaciones.

3.4. Correlaciones parciales y semiparciales. La correlación parcial se trata del remanente que hay entre las variables después de haber eliminado la correlación debida a su asociación mutua con otras variables, mientras que la correlación semiparcial es la que hay entre la variable dependiente e independiente, una vez que se ha eliminado de la variable independiente los efectos lineales de las otras variables independientes.

3.5. Diagnósticos de colinealidad. Verifica que la variable independiente no sea una función lineal de otra variable, es decir, que sean completamente independientes.

El botón de “Gráficos” nos permite obtener los gráficos de dispersión, y el histograma. En la lista de variables tenemos:

- DEPENDNT. Variable dependiente.
- *ZPRED. Valores pronosticados tipificados.
- *ZRESID. Residuos tipificados.
- *DRESID. Residuos eliminados.
- *ADJPRED. Valores pronosticados corregidos.
- *SRESID. Residuos estudentizados.
- *SDRESID. Residuos estudentizados eliminados.

Tal vez sea extraña la razón por la cual no aparecen las variables originales. La razón es que gracias a estos cambios podremos detectar valores atípicos, observaciones poco usuales y casos influyentes.

Con el botón Guardar tendremos los valores de la lista de variables que mencionamos en la opción de “Gráficos”; las categorías son las siguientes:

- Valores pronosticados: Son los valores pronosticados usando la regresión lineal.
- Distancias: Se trata de medidas para identificar casos con combinaciones poco usuales de valores para las variables independientes y casos que puedan tener un gran impacto en el modelo.
- Intervalos de pronóstico: Proporciona los límites superior e inferior para los intervalos de pronóstico individual y promedio.
- Residuos: Es la diferencia del valor observado de la variable con respecto al valor pronosticado.
- Estadísticos de influencia: Involucra los cambios en los coeficientes de regresión (DfBetas), valores pronosticados (DfAjuste), así como sus tipificaciones.
- Estadísticos de los coeficientes: Almacena los coeficientes de la regresión en un conjunto de datos.

Con respecto al botón “Opciones”, nos brinda las siguientes:

- Criterios del método por pasos. Recomendables cuando el método de selección es diferente de Introducir y pasos sucesivos.
- Incluir la constante en la ecuación. Se refiere a la ordenada a la origen, si lo deseamos podemos eliminarla de nuestro modelo de regresión.

- Valores perdidos. Esta sección es la correspondiente para el tratamiento de los valores perdidos.

Con la última sección, si lo deseamos, usaremos una variable de elección para limitar el análisis al conjunto de datos que cumpla con la “Regla”. Con etiquetas de caso se identificarán los puntos en los diagramas. Por último, la Ponderación MCP es la ideal cuando se usa mínimos cuadrados.

Hemos comentado cómo usar la regresión, sin embargo, en muchas ocasiones puede no ser la mejor opción. Por el momento no hablaremos de cómo identificar este punto, ya que resulta más práctico verlo en el ejemplo general.

4.2.2 REGRESIÓN LINEAL MÚLTIPLE

En la práctica hay análisis de regresión en los que necesitamos más de una variable independiente dentro del modelo, estos son los modelos de regresión múltiple. El modelo tendrá k variables independientes y k coeficientes, la forma que tendrá es:

$$\hat{Y} = B_0 + B_1x_1 + B_2x_2 + \dots + B_kx_k$$

La representación gráfica de esta regresión no es una línea recta, esto es porque están involucradas más de dos variables dentro de la relación, de tal forma que tendremos un plano, y cada variable independiente corresponderá a un factor.

SPSS no proporciona una categoría especial para la regresión lineal múltiple, sin embargo, el cuadro de diálogo de “Regresión lineal” nos servirá para construir los modelos de regresión lineal múltiple, bastará con depositar más de una variable en la casilla de “Independientes”.

La primer tabla que SPSS arroja como resultado contiene las variables involucradas en el modelo y el método usado; el resto de las tablas corresponden a las mismas que se obtienen en la regresión lineal simple.

4.3 EJEMPLO GENERAL

En los ejemplos anteriores limitábamos los estudios a cierta región, país, sector, etc., sin embargo, ¿no resultaría interesante analizar algún tema a nivel mundial? ¿Pero cómo? Pues bien, con GapMinder podremos hacerlo, a manera de introducción podemos leer el documento de la página web: <http://issuu.com/elsoftwarevolandero/docs/31-gapminder/1>.

Una vez que ingresemos a la página de Gapminder, buscaremos dentro de “Data”, en la página 3, el nombre del indicador “Cell pones (total)” y descargaremos el archivo de Excel.

Los datos del indicador “Cell pones (total)” se refieren al número de suscriptores que cuentan con acceso a la red telefónica mediante teléfono celular, incluyendo sistema de prepago y pospago.

Para este ejemplo, nuestra hipótesis es que la tendencia del número de clientes a lo largo del tiempo fue en aumento, a nivel mundial.

La información que contiene está por año y país, así que debido a la forma en que está será necesario hacerle los cambios siguientes:

1. Eliminar las columnas inferiores al año 1980.
2. Crear una nueva hoja con el nombre de “Datos”.
3. Seleccionar todos los nombres de los países.

4. Hacer un pegado especial en la hoja “Datos”, para ello daremos clic derecho, luego seleccionar pegado especial y por último “Trasponer”. De modo que cada país será una variable.
5. Al principio de los nombres de los países tendremos la variable “Año”, para ello repetiremos los pasos 3 y 4 con los años, no con los valores.
6. Seleccionar y pegar de forma transpuesta el resto de los valores en nuestra hoja. Como resultado tendremos 218 variables del año 1980 a 2008.

El siguiente paso es el más sencillo de todos, se trata de pasar nuestros datos del archivo de Excel a SPSS.

Dado que son demasiadas variables, sería más difícil analizarlas, así que elegiremos las siguientes: Afganistán, Argentina, Bulgaria, China, Dominica, Ecuador, France, Gabón, Italy y Japan.

Tal como dijimos en los tips, antes de hacer una correlación veamos los gráficos de caja, y de dispersión. En primera instancia observemos el gráfico de caja, tienen en su mayoría datos atípicos debido a que no todos los países empezaron a tener suscriptores en el mismo año.

Con respecto a los diagramas de dispersión, el de Afganistán y Argentina tienen un comportamiento lineal hasta el año 2005, pero en general todas las tendencias del número de suscriptores a través del tiempo son exponenciales, lo cual es bastante lógico debido a su creciente uso.

Ahora chequeemos cómo están las correlaciones de China, Bulgaria y Ecuador con respecto al tiempo, es decir, qué tanto explica la variable del tiempo al número de suscriptores del servicio de telefonía celular.

En la tabla 19 tenemos el resumen de la correlación que habíamos mencionado en el párrafo anterior. Cada variable tuvo un total de 29 casos, además de que a cada relación SPSS colocó doble asterisco, lo cual indica que la correlación es significativa al nivel 0.01 bilateral.

	Correlación de Pearson	Sig. (bilateral)
Año	1.000	.000
Bulgaria	.745	.000
China	.788	.000
Ecuador	.705	.000

Tabla 19. Correlación de Pearson

El siguiente paso será obtener nuestros coeficientes para nuestra regresión lineal.

La tabla 20 es la primera que nos arrojó SPSS como resultado de la regresión lineal. El contenido de esta se refiere al coeficiente de Pearson. El valor de R cuadrado fue ligeramente alto, lo cual significa que la variable Año explica al número de suscriptores en un 55.6%.

En cuanto al valor de R corregida no varía mucho de R cuadrada debido a que sólo tenemos una variable independiente. El *error típico de estimación* se refiere a la desviación típica de los residuos resultantes entre los valores establecidos y los obtenidos por el modelo de regresión lineal, así que, cuanto menor sea mayor es el ajuste.

Sigamos entonces, con la tabla ANOVA sabremos si existe relación significativa entre las variables. La idea de la prueba es contrastar el valor de R , si este vale cero entonces la pendiente de la recta de regresión también es cero. Con respecto a nuestro modelo, las variables están linealmente relacionadas.

		Modelo
		1
R		.745
R cuadrado		.556
R cuadrado corregida		.539
Error típ. de la estimación		2149142.150
	Cambio en R cuadrado	.556
Estadísticos de cambio	Cambio en F	33.766
	gl1	1
	gl2	27
	Sig. Cambio en F	.000

Tabla 20. Resumen del modelo

Modelo	Suma de cuadrados	gl	Media cuadrática	F	Sig.
1 Regresión	155960885016882.000	1	155960885016882.00	33.766	0.00
Residual	124707923465964.000	27	4618811980220.90		
Total	280668808482846.000	28			

Tabla 21. ANOVA

Hasta el momento no tenemos idea de cómo es nuestro modelo de regresión lineal, pero sabemos cuál es la estructura que debe seguir, lo que necesitamos son los coeficientes. En la tabla 22 tenemos los *Coefficientes*, la constante será nuestra B_0 , y el valor debajo de esta corresponde a B_1 y estará asociada con la variable *Año*.

Modelo	Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
	B	Error típ.	Beta		
1 (Constante)	-551013354.076	95114362.141		-5.793	0.00
Año	277178.683	47699.862	.745	5.811	0.00

Tabla 22. Coeficientes

De esta manera, nuestro modelo queda la forma siguiente: $\text{Suscriptores} = 277178.683 \text{ Año} - 551013354.076$

Con la ecuación anterior bastará con el valor de un año y tendremos el número aproximado de suscriptores, de esa fecha en cuestión, para Bulgaria.

Por último, la sección de los valores de t y de $Sig.$ nos ayudarán a determinar si un coeficiente de regresión es significativamente distinto de cero, en este caso ambos coeficientes lo son.

El proceso para obtener los otros modelos que mencionamos para los países de China, y Ecuador es el mismo, de tal forma que las ecuaciones resultantes serán:

$$\text{China: Suscriptores} = .00000017 \text{ Año} - 00000000344$$

Ecuador: Suscriptores= 267951.290 Año-000000532

En el caso de China, la variable Año explica al número de suscriptores en un 60.7%, y en un 47.8% a Ecuador.

En general vimos que la tendencia de este “fenómeno” telefónico llegará a un punto de equilibrio, de ese modo, también podemos observar que esta afluencia de clientes puede deberse también a la oferta tanto de los servicios así como de la innovación de la tecnología celular. En conclusión nuestra hipótesis no fue correcta, ya que si bien ha ido aumentando el número de suscriptores, tendremos un punto en que dejará de crecer para establecerse en un punto de equilibrio.

4.4 EJERCICIO COMPLEMENTARIO

En éste ejercicio ocuparemos los datos del estudio llamado “Light-Duty diesel emission correction factors for ambient conditions”. Consiste en saber si la emisión de óxido nitroso de un camión de reparto ligero a diesel tiene relación con los niveles de humedad, temperatura del aire y presión barométrica.

Cada medición de la emisión fue tomada en diferentes momentos del tiempo. Lo que se desea es establecer un modelo de regresión lineal múltiple, en donde el óxido nitroso será nuestra variable dependiente, y el resto de los factores corresponden a las variables independientes del modelo.

En este ejercicio, nuestra hipótesis es que el modelo de regresión lineal múltiple es el adecuado, ya que por tratarse de una reacción química puede tener afectaciones con cuestiones ambientales tales como la humedad, la temperatura y la presión.

Los datos que ocuparemos los podemos obtener del libro “Probabilidad y estadística para ingenieros”, del autor Ronald E. Walpole, pero qué lata tener que pasar los datos, así que si los quieres tener en un archivo puedes mandar un mail a gonzalez_maldonado_ec@yahoo.com.mx, coloca en el asunto “Light-Duty” y tendrás los datos en un archivo de Excel.

El siguiente paso es pasar nuestros datos a SPSS. No olvidemos que antes de obtener la regresión lineal múltiple es preferible usar la correlación, en este caso la parcial.

Como resultado tenemos que la correlación que existe entre estas variables es negativa (ver Tabla 23), así que ahora veamos qué pasa con el modelo de regresión lineal múltiple. Para ello colocaremos en Dependientes la variable del Óxido Nitroso y en la casilla de dependientes colocaremos el resto de las variables.

Con la Tabla 24 vemos que la humedad, la temperatura y la presión explican en un 76.3% a la emisión de óxido nitroso, que es un valor bastante bueno, así que ahora toca revisar qué pasa con respecto a los valores de los coeficientes y el modelo en general.

La tabla del ANOVA tiene un valor de significancia de cero, por lo que el conjunto de variables independientes ofrecen un buen ajuste para la variable dependiente.

Usando la tabla de Coeficientes, nuestro modelo de regresión lineal múltiple queda de la siguiente forma:

$$\text{Óxido Nitroso} = .154 \text{ Presión} + 0.001 \text{ Temperatura} - 0.003 \text{ Humedad} - 3.508$$

La ecuación anterior es exactamente la misma que en el libro de Walpole, lo que indica que los resultados de SPSS son bastante eficientes.

La interpretación que debemos darle a los coeficientes no es propiamente independiente, ya que los valores de cada uno de ellos fueron obtenidos tomando en cuenta el resto de los valores de las otras variables.

Variables de control			Humedad	Temperatura	Presión
Óxido Nitroso	Humedad	Correlación	1.000	.201	-.251
		Significación (bilateral)	.	.409	.297
		gl	0	17	17
	Temperatura	Correlación	.201	1.000	-.086
		Significación (bilateral)	.409	.	.725
		gl	17	0	17
	Presión	Correlación	-.251	-.086	1.000
		Significación (bilateral)	.299	.725	.
		gl	17	17	0

Tabla 23. Correlaciones parciales

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	.895	.800	.763	.05617

Tabla 24. Resumen del modelo múltiple

Modelo	Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
	B	Error típ.	Beta		
1 (Constante)	-3.508	3.005		-1.167	.260
Humedad	-.003	.001	-.693	-4.008	.001
Temperatura	.001	.002	.045	.391	.701
Presión	.154	.101	.259	1.521	.148

Tabla 25. Coeficientes múltiple

4.5 ¿OTRO EJERCICIO COMPLEMENTARIO? POR QUÉ NO...

Hemos visto un ejemplo de regresión lineal simple y otro de múltiple, sin embargo, podemos encontrarnos en la situación de elegir cuál podemos modelar, es decir, que debido a la naturaleza de los datos es posible usar uno u otro tipo de regresión.

La Tabla 26 contiene los datos de 12 personas, correspondientes al peso, edad y estatura, las cuales están medidas en libras, pulgadas y años respectivamente.

Peso	64	71	53	67	55	58	77	57	56	51	76	68
Estatura	57	59	49	62	51	50	55	48	52	42	61	57
Edad	8	10	6	11	8	7	10	9	10	6	12	9

Tabla 26. Peso, estatura, edad

Si trabajáramos para una institución de salud, ¿cuál sería nuestra primera relación que estableceríamos? Posiblemente la del peso con la estatura, o tal vez la edad con el peso. Tomando en cuenta lo anterior veamos qué pasa con la correlación que existe entre estas variables, para ello usaremos la correlación bivariada.

Como resultado tenemos que la relación es más fuerte entre la variable Peso y Estatura, mientras que la estatura y la edad es la segunda con el coeficiente más alto.

Hagamos ahora dos modelos de regresión lineal, en ambos la variable dependiente será la Estatura. Las variables independientes corresponderán al Peso y a la Estatura respectivamente. No es recomendable usar las capas, ya que a partir de la segunda capa se toma en cuenta a las variables independientes de las anteriores, generando entonces un modelo de regresión lineal múltiple.

Con la tabla 27, vemos que el modelo 1 explica en un 63.9%, mientras que el modelo 2 es relativamente menor. Usando el ANOVA podemos decir que la ecuación de regresión lineal definida por los modelos 1 y 2 ofrece un buen ajuste para los datos.

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	.820	.672	.639	3.573
2	.798	.637	.601	3.755

Tabla 27. Resumen modelos

Modelo	Suma de cuadrados	gl	Media cuadrática	F	Sig.
1 Regresión	261.281	1	261.281	20.471	.001
Residual	127.635	10	12.764		
Total	388.917	11			
2 Regresión	288.044	2	144.022	12.850	.002
Residual	100.873	9	11.208		
Total	388.917	11			

Tabla 28. ANOVA modelos

Por el momento quedémonos con el modelo 1:

$$\text{Estatura} = 19.55 + 0.542 \text{ Peso}$$

La siguiente alternativa es establecer un modelo que involucre tanto a la Edad como al Peso.

El porcentaje de explicación es de 68.3, lo que significa que tiene un mayor ajuste que el modelo que implica únicamente al peso. De acuerdo a los coeficientes que nos arrojó SPSS, la ecuación de regresión múltiple para la Estatura queda la forma siguiente (al que llamaremos modelo 3):

$$\text{Estatura} = 21.320 + 0.333 \text{ Peso} + 1.287 \text{ Edad}$$

Una vez que contamos con nuestras regresiones, toca ver cómo aplicarlas. Recordando la forma de calcular una nueva variable, bastará con introducir los valores de la ecuación y asociarles la respectiva variable, de esta forma tendremos como resultado la estatura pero a partir de la regresión que construimos.

Llamaremos a las nuevas variables `est_reg` (con el modelo 3), y `est_reg2` (con el modelo 1); en seguida SPSS creará las dos variables, y a primera vista pareciera que los valores del `est_reg2` se acercan más a los valores reales de la Estatura, sin embargo, hay datos en los que éste supera por más de una unidad.

En conclusión, no es posible generalizar el modelo de regresión múltiple para todo experimento que deseemos modelar, ya que éste siempre dependerá de la información con la que contemos. Por ejemplo, si el coeficiente de correlación de la Edad con el Peso hubiera sido de un 0.84, entonces lo recomendable es usar un modelo que involucre a estas variables, sin embargo, esta decisión no sólo dependerá de lo que se pueda relacionar sino de qué queremos modelar.

4.6 CASO DE ESTUDIO (CONTINUACIÓN...)

Si bien los datos reflejan una cosa, contar con la opinión del sujeto estudiado comprende una herramienta más para el analista; por lo que ahora toca revisar las respuestas de la encuesta.

Por lo que se refiere al grupo que no usó SPSS, la mayoría no tenía gusto por la estadística, razón que podemos vincular con la necesidad de ésta en el trabajo, así como haber tomado un curso dentro de la licenciatura cursada. Tomando en cuenta lo anterior, sólo el 50%(3) usa algún análisis para sus actividades laborales, mientras que el 83.33% (5) opinaron que no quedaron satisfechos en términos de lo aprendido relacionado con la estadística, en este mismo sentido el 100% (6) estuvo de acuerdo en que en su aprendizaje de la estadística le hubiera gustado que se utilizara algún tipo de software.

Veamos ahora qué comentó el grupo que usó SPSS. El 50% (3) aceptó tener gusto por la estadística, mientras que el resto respondió que no es de sus materias preferidas. Además, el 100% no utiliza con frecuencia la estadística, sin embargo, ese mismo 100% estuvo satisfecho en un 95% de lo aprendido en el curso que tomó durante su licenciatura.

Tomando en cuenta lo anterior, también afirmaron en un 100% que sí les hubiera interesado el manejo de software, específicamente para agilizar la obtención de estadísticos. Asimismo, la mayoría de los integrantes estuvieron de acuerdo en que tanto el uso de un ejemplo real así como el manejo de software facilitaron el aprendizaje del tema planteado.

En un mundo académico ideal, el factor del software significaría un mejoramiento uniforme en el aprendizaje de la estadística, sin embargo, la realidad es que no sólo se ve involucrado la herramienta de software usada, sino que es un conjunto de aspectos que pueden ser trabajados, tales como el gusto por el objeto de estudio, la necesidad de usar lo aprendido, los conocimientos previos; todos ellos mencionados por los integrantes de los equipos durante el desarrollo del experimento

De acuerdo con la prueba de hipótesis no hubo diferencia significativa entre ambos métodos, pero esto no significa que siempre vaya a ser así, ya que faltaría realizar una combinación de los métodos y de los puntos que menciona Behar [5].

Además, se puede pensar en que el software no importa, sin embargo, más que usar un software en específico debe tenerse en mente que es una herramienta de apoyo para la obtención de estadísticas, ya que el análisis dependerá de la experiencia y conocimientos de cada persona.

Capítulo 5. SERIES DE TIEMPO

Keep It Simple Stupid.
Principio KISS

Objetivo: En este capítulo conoceremos distintas metodologías que podemos aplicar para hacer pronósticos de una serie de tiempo.

Las series de tiempo son un poderoso instrumento para la toma de decisiones, con base en su análisis y modelado surgen los pronósticos, y a partir de ellos podemos plantear distintos escenarios según nos sea conveniente.

Las series de tiempo están compuestas por un conjunto de observaciones (que pueden ser de tipo continuo o discreto) ocurridas en un espacio de tiempo (variable discreta).

Dado que usaremos una variable que involucra el tiempo debemos configurarla para establecer si se trata de datos medidos semanalmente, anualmente, etc. o incluso si son mediciones en horas, minutos. En el menú Datos → Definir fechas, podremos elegir la frecuencia con que están medidos nuestros datos (ver Fig. 46).



Fig. 46. Definir fechas

Hablemos un poco acerca de los elementos que están involucrados en una serie de tiempo.

1. **Tendencia.** Se refiere al comportamiento que tendrán nuestros datos. Puede tratarse de un crecimiento exponencial o lineal, o de una curva que crezca y luego se encuentre en un punto de equilibrio. (Ver Fig. 47a)
2. **Variación estacional.** Tiene que ver con la repetición de patrones en un período menor o igual a un año, es por ello que lleva el apellido “estacional”. Por ejemplo, el número de vacacionistas aumenta en época de vacaciones, las ventas de una tienda de regalos. (ver Fig. 47b)
3. **Ciclo.** En este caso la longitud de la variación estacional será mayor a un año; suele ser más difícil de detectar debido a que para recolectar los datos se requiere de más tiempo. El ejemplo más común es el ciclo económico, el aprovechamiento agrícola. (ver Fig. 47c)

4. Fluctuación aleatoria. Son variaciones irregulares de nuestra serie, es decir, situaciones que no son explicadas con el modelo matemático. Por ejemplo, que en los períodos vacacionales haya ocurrido un fenómeno meteorológico, o que el sindicato de pilotos de avión se haya puesto en huelga en una temporada en que el servicio de vuelos en avión es muy demandado.

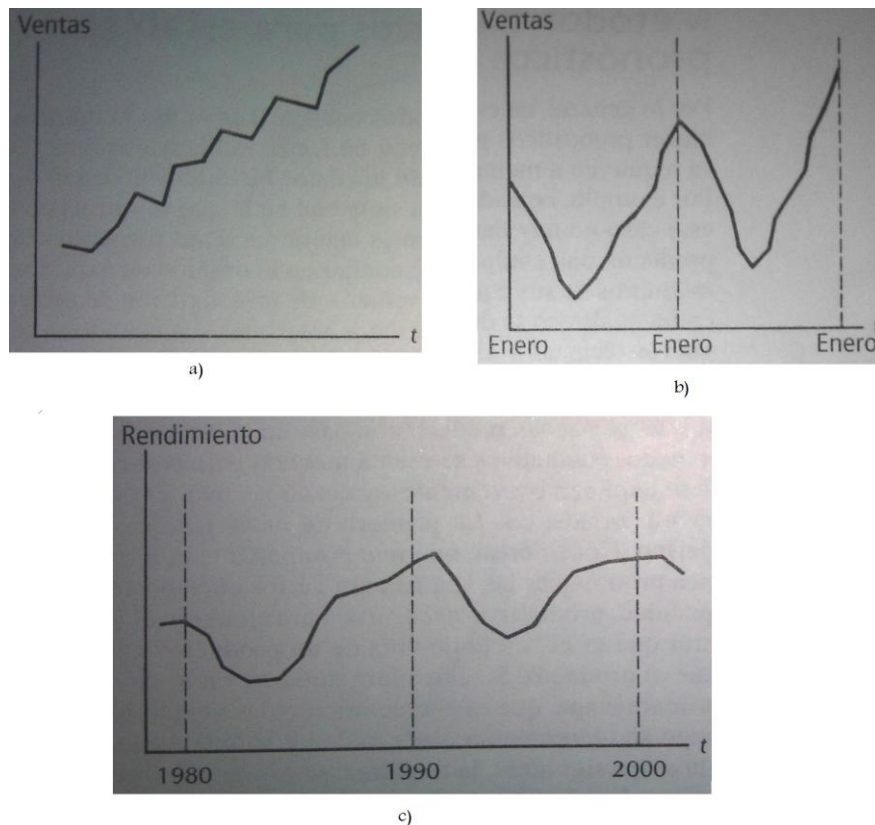


Fig. 47. Ejemplos series de tiempo (FUENTE: Bowerman, 2007)

La forma de visualizar los elementos mencionados es con la gráfica de la serie de tiempo. La obtendremos en Analizar → Predicciones → Gráficos de secuencia (ver Fig. 48), bastará con colocar a la variable de las observaciones en la casilla de variables, y en “Etiquetas del eje de tiempo” la variable que corresponda a la medición del tiempo.

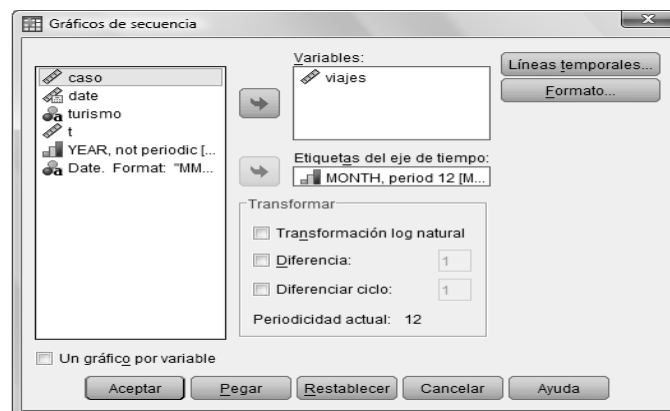


Fig. 48. Gráficos de secuencia

En las gráficas de la Fig. 47, tenemos reflejados los elementos uno a tres que se pueden presentar en una serie de tiempo, sin embargo, es posible que dentro de un mismo conjunto de datos detectemos uno o más

de éstos elementos, para ello veamos ahora la gráfica de la Fig. 49, la cual corresponde al número viajes realizados cada mes, en donde podemos ubicar tendencia, variación estacional y fluctuación aleatoria.

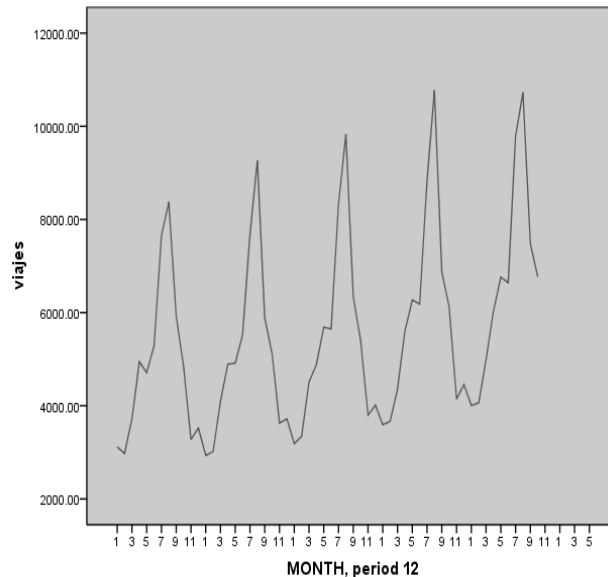


Fig. 49. Serie Viajes

Continuemos, para la elaboración de modelos de pronósticos contamos con varias herramientas de ayuda, mismas que veremos en los siguientes temas. Antes de continuar, uno de los términos muy usado dentro de los pronósticos es el error que existe entre el valor verdadero y el valor obtenido con el modelo de pronóstico, el cual está definido por:

$$e_t = Y_t - \hat{Y}_t$$

En donde:

\hat{Y}_t : Valor pronosticado de Y_t

Y_t : Valor real de la variable Y en el tiempo t

Cabe señalar, que esta es una de la formas en que se puede medir el error en que se incurre cuando realizamos pronósticos, para mayor información podemos consultar acerca de los errores cuadráticos y el cuadrático medio.

5.1 FUNCIÓN DE AUTOCORRELACIÓN Y DE AUTOCORRELACIÓN PARCIAL

La función de autocorrelación nos sirve para saber el efecto que tiene la variación de un dato en otra etapa, y se encuentra dentro del menú Analizar → Predicciones → Autocorrelaciones (ver Fig. 50).

A diferencia de la función de autocorrelación, la función de autocorrelación parcial tiene por objetivo mostrar qué tan importante es conservar una variable dentro del modelo. La forma de obtenerse mediante el programa es la misma que para la autocorrelación.

La función de autocorrelación estará dada por la fórmula siguiente:

$$\begin{aligned}\rho_k &= \text{Corr}(Y_t, Y_{t+k}) \\ &= \frac{\text{Cov}(Y_t, Y_{t+k})}{\sqrt{\text{Var}(Y_t)\text{Var}(Y_{t+k})}} \\ &= \frac{\gamma_k}{\gamma_0}\end{aligned}$$

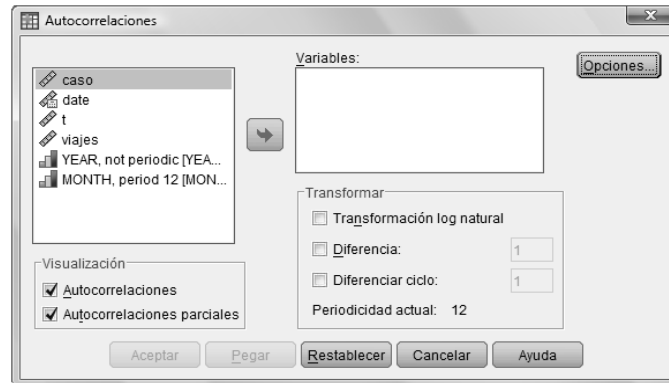


Fig. 50. Autocorrelaciones

5.2 PERIODOGRAMA

Además de los gráficos de ambas autocorrelaciones, contamos también con el periodograma el cual permite saber si existe periodicidad en la serie de tiempo. Este tipo de gráfico consiste en la exageración de la amplitud para facilitar la detección de la periodicidad en los datos si es que hubiere.

En el menú Analizar → Predicciones → Análisis espectral, se puede acceder al cuadro de diálogo de la Fig. 51, con el que obtenemos el periodograma. Es muy importante que los datos no contengan valores perdidos por el sistema.



Fig. 51. Periodograma

5.3 MODELOS ESTACIONARIOS Y NO ESTACIONARIOS

Antes de que empecemos a conocer los modelos para las series de tiempo es importante verificar si dicha serie es estacionaria en sentido amplio, es decir:

1. Su función de media sea constante: $E(Y_t) = E(Y_{t+r}) = \mu, \quad \forall r \in \mathcal{R}$

2. La varianza sea constante: $Var(Y_t) = Var(Y_{t+r}) = \gamma_0, \quad \forall r \in \mathfrak{R}$
3. La función de autocorrelación no dependa de la posición del tiempo: $Autocorr(Y_t, Y_{t+r}) = Autocorr(Y_{t+r}, Y_{t+r-k}) = \rho_k, \quad \forall r \in \mathfrak{R}$

Si la serie de tiempo no cumple con alguna propiedad de estacionaridad podemos realizar algunas de las siguientes opciones dependiendo de cual se trate:

1. Para estabilizar la varianza podemos aplicar alguna transformación matemática como tal como la función del logaritmo natural, raíz cuadrada, recíproco.
2. Eliminación de la tendencia, para ello podemos realizar diferencias ordinarias (también llamadas finitas o no estacionales), las cuales consisten en restar los valores de las variables con respecto al tiempo, de este modo, por cada diferencia ordinaria perderemos una observación. A continuación veamos las diferencias de grado 1 a 3.

$$\begin{aligned}\Delta Y_t &= Y_t - Y_{t-1} \\ \Delta^2 Y_t &= \Delta Y_t - \Delta Y_{t-1} \\ &= (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) \\ &= Y_t - 2Y_{t-1} + Y_{t-2} \\ \Delta^3 Y_t &= Y_t - 3Y_{t-1} + 3Y_{t-2} - Y_{t-3}\end{aligned}$$

3. Variación estacional, la cual podemos quitar con diferencias estacionales. Éstas las obtendremos de la manera siguiente:

$$\begin{aligned}\Delta_s Y_t &= Y_t - Y_{t-s} \\ \Delta_s^2 Y_t &= \Delta_s Y_t - \Delta_s Y_{t-s} \\ &= (Y_t - Y_{t-s}) - (Y_{t-s} - Y_{t-2s}) \\ &= Y_t - 2Y_{t-s} + Y_{t-2s} \\ \Delta_s^3 Y_t &= Y_t - 3Y_{t-s} + 3Y_{t-2s} - Y_{t-3s}\end{aligned}$$

El valor de s corresponde a la longitud del período de la variación estacional. Supongamos que en la serie de tiempo encontramos un comportamiento que se repite cada cuatro meses, entonces el valor de s será de 3; en cambio si observamos una repetición 12 meses, entonces s valdrá 12. En contraste con las diferencias ordinarias, por cada diferencia estacional perderemos s observaciones.

Cualquiera de estas opciones las podemos obtener mediante los cuadros de diálogo de los gráficos de secuencia, de las autocorrelaciones o en el de Diagramas espectrales. Cabe mencionar que si la serie de tiempo no cumple con ninguna de las propiedades, entonces se le debe dar tratamiento primero a la estabilización de la varianza, de este modo se evita obtener datos negativos derivados de alguna transformación matemática.

5.3.1 MODELOS AUTOREGRESIVOS Y DE MEDIAS MÓVILES

Los primeros modelos que conoceremos son los procesos autoregresivos, que forman parte del modelo lineal general. Se denotan con las letras **AR (p)**, están basados en la historia, es decir, los valores que conforman a la serie de tiempo.

El modelo matemático es el siguiente:

$$Y_t = \delta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \dots + \phi_p Y_{t-p} + e_t$$

En donde:

$$\delta = \mu(1 - \phi_1 - \phi_2 - \dots - \phi_p).$$

$\phi_1, \phi_2, \phi_3, \dots, \phi_p$ son parámetros desconocidos que relaciona a los valores de $Y_{t-1} + Y_{t-2} + Y_{t-3} + \dots + Y_{t-p}$.

e_t : es el error aleatorio.

Por ejemplo, el modelo AR (1) es:

$$Y_t = \delta + \phi_1 Y_{t-1} + e_t$$

Mientras que el AR (4) corresponde a:

$$Y_t = \delta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \phi_4 Y_{t-4} + e_t$$

La forma de obtener estos modelos en SPSS, la veremos hasta el tema de modelos ARIMA.

Los modelos de medias móviles están basados en la ponderación de los errores aleatorios, de modo que no toma en cuenta la historia, se denota por **MA (q)**.

El modelo matemático es el siguiente:

$$Y_t = \delta + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \theta_3 e_{t-3} - \dots - \theta_q e_{t-q}$$

En donde:

$$\delta = \mu$$

$e_t, e_{t-1}, e_{t-2}, e_{t-3}, \dots, e_{t-q}$: son los errores aleatorios en el momento t.

$\theta_1, \theta_2, \theta_3, \dots, \theta_q$: son los parámetros que se relacionan con los errores aleatorios.

Por ejemplo, el modelo MA (3) es:

$$Y_t = \delta + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \theta_3 e_{t-3}$$

Mientras que un modelo MA (1) tiene la forma:

$$Y_t = \delta + e_t - \theta_1 e_{t-1}$$

Podremos preguntarnos, ¿es posible combinar los modelos anteriores? La respuesta es afirmativa, dando origen a los modelos **ARMA (p,q)**, los cuales contendrán p+q parámetros, con el modelo matemático a continuación mostrado:

$$Y_t = \delta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \dots - \theta_q e_{t-q}$$

Por ejemplo, el modelo ARMA(3,2) nos queda como:

$$Y_t = \delta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}$$

5.3.2 MODELOS ARIMA

En el tema anterior vimos que existen modelos: **AR (p)**, **MA (q)** y **ARMA (p, q)**, los cuales están basados ya sea en los valores de la serie de tiempo o en los errores aleatorios; sin embargo, como ya habíamos mencionado al inicio del tema, pudimos haber aplicado alguna transformación de logaritmo natural, después aplicar dos diferencias ordinarias, y finalmente una diferencia estacional; por lo anterior los datos ya no son con los que contábamos al principio.

De acuerdo a lo anterior, ahora nuestro modelo matemático es:

$$W_t = \delta + \phi_1 W_{t-1} + \dots + \phi_p W_{t-p} + e_t - \theta_1 e_{t-1} + \dots - \theta_q e_{t-q}$$

Éste corresponde a un **ARIMA (p, d, q)**, que contempla la parte AR(p), MA(q) y alguna transformación para que la serie sea estacionaria (I).

Dentro de la opción “Crear modelos”, encontraremos el modelizador de series temporales (ver Fig. 52), el cual cuenta con los modelos:

- Modelizador experto. SPSS crea un modelo para los datos.
- Suavizado exponencial. Este tipo de modelo lo veremos en el tema 5.4.
- ARIMA.

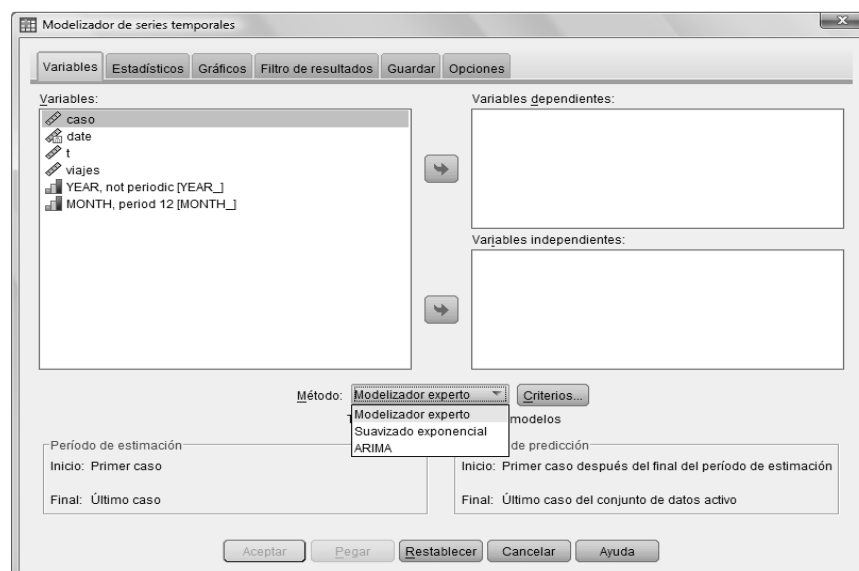


Fig. 52. Modelizador de series temporales

Veamos ahora cómo configurar el modelo ARIMA, gráficos y estadísticos, entre otras opciones.

En la casilla de "Variables dependientes" debemos colocar la variable que determina la observación, y en la de "Variables independientes" pondremos la variable que mide el tiempo.

En la esquina inferior izquierda tenemos el período de estimación, por default es de la fecha en que comienzan y terminan, se refiere al punto en que se tomarán los datos para la creación del modelo.

En contraparte, el período de predicción corresponderá a los datos que serán pronosticados, cuenta con valores predeterminados; sin embargo, es posible cambiarlos.

La ficha de estadísticos tiene tres secciones:

- Medidas de ajuste. Tiene que ver con los valores que nos ayudarán a decidir qué tan bien se ajusta el modelo ARIMA a los datos.
- Estadísticos de comparación de modelos. Nos servirán para comparar el modelo y los datos.
- Estadísticos de modelos individuales. Tal como su nombre lo dice los estadísticos para cada modelo.

Dentro de la pestaña de gráficos contamos con dos secciones:

- Gráficos para comparar modelos
- Gráficos de modelos individuales

Ambos tienen que ver con los estadísticos, sólo que en esta ocasión se trata de la gráfica de estos.

El filtro de resultado funciona de la misma manera que "Seleccionar casos".

Para que SPSS no solamente nos haga el modelo sino que también nos guarde los valores que se están obteniendo debemos configurarlo dentro de la ficha "Guardar".

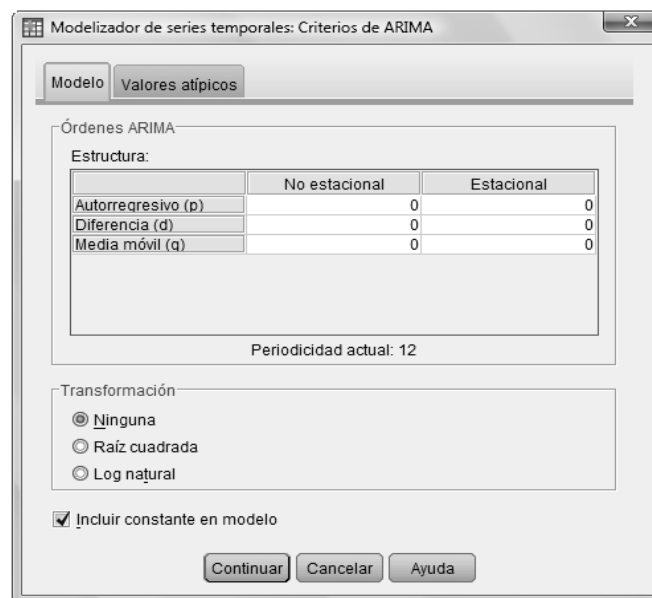


Fig. 53. Criterios de ARIMA

Por último dentro de la ficha “Opciones” controlaremos el “Período de predicción”, bastante útil, ya que con este definiremos nuestro pronóstico.

Hasta el momento hemos visto lo relacionado a configurar un modelo ARIMA (p, d, q), sin embargo, si lo que necesitamos es un AR (3), MA (4), o un ARIMA (2, 1, 3) debemos presionar el botón “Criterios” (ver Fig. 53).

En la primera parte definiremos el orden del modelo, lo que determinará si se trata de un AR (3) o un MA (4); o en su caso si es necesario hacer alguna transformación y si incluiremos la constante al modelo.

No olvidemos que los modelos ARIMA requieren de al menos cierta cantidad de datos.

5.4 SUAVIZAMIENTO EXPONENCIAL

Otro de los métodos para construir modelos de pronóstico es el suavizamiento exponencial. Éste consiste en hacer los datos más parecidos a una curva suave, para ello se requiere que la varianza sea constante.

La esencia del suavizamiento exponencial es darle una mayor ponderación a las observaciones más recientes, esto es, mediante una o más constantes de suavización.

Existen varios métodos que usan el suavizamiento, SPSS los divide en estacionales y no estacionales.

El proceso para obtener estos modelos es el mismo que para el método ARIMA, así que ahora veamos en qué consisten estos métodos.

5.4.1 MODELOS NO ESTACIONALES

Una vez que hayamos elegido en el Modelizador el método de “Suavizamiento exponencial”, para definir de qué tipo se trata, debemos dar clic en Criterios.

Dentro de la sección “Tipo de modelo”, tenemos los modelos no estacionales que maneja, los cuales describiremos a continuación:

1. Simple. Se trata del caso más sencillo de suavizamiento exponencial, éste modelo es el adecuado cuando en la serie temporal no hay variación estacional, pero la media cambia lentamente en el tiempo. Consiste en la suma de ponderaciones de los valores anteriores.

$$\hat{Y}_{t+1} = \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + \alpha(1-\alpha)^2 Y_{t-2} + \dots + \alpha(1-\alpha)^{N-1} Y_{t-(N-1)}$$

El valor de α es la constante de suavización entre 0 y 1.

2. Tendencia lineal de Holt. Esta es una de las vertientes del suavizamiento exponencial doble, realiza un suavizamiento de las constantes de suavización por separado.
3. Tendencia lineal de Brown. Aplica la fórmula del suavizamiento doble, de modo que encuentre el valor óptimo de alpha, ya que minimiza la suma de cuadrados de los residuales.

$$\hat{Y}_{N+h} = \hat{\beta}_{0,N} + \hat{\beta}_{1,N} h$$

4. Tendencia amortiguada.

La particularidad de estos modelos es que son no estacionales, es decir, que no tienen variación estacional. Para cada uno de ellos podremos configurar los estadísticos, variables, etc.

5.4.2 MODELOS ESTACIONALES

Habr  ocasiones en las cuales a pesar de las transformaciones matem ticas o de las diferencias (ordinarias o estacionales) que hagamos a la serie de tiempo no sea posible eliminar la variaci n estacional, por lo que ser  necesario modelarla.

SPSS cuenta con los modelos estacionales a continuaci n descritos:

1. Estacional simple.
2. Aditivo de Winters. Tal como lo dice su nombre, se trata de la suma de los efectos de la variaci n estacional, es posible usarlo cuando la variaci n estacional es constante con respecto a la tendencia, es decir, que la variaci n estacional no se incrementa con respecto a la tendencia.
3. Multiplicativo de Winters. En este caso si las variaciones estacionales son proporcionales al nivel pero los errores son aditivos, entonces podemos usar este m todo.

Por  ltimo, del lado derecho tenemos una secci n para aplicarle alguna transformaci n matem tica a nuestra serie de tiempo; las m s comunes son el logaritmo y la ra z cuadrada.

5.5 DESCOMPOSICI N ESTACIONAL

Hasta el momento hemos visto los distintos tipos de modelos que podemos utilizar para nuestra serie de tiempo, sin embargo, conforme vamos avanzando es necesario usar t cnicas m s finas para que nuestros pron sticos tengan un buen ajuste.

A diferencia de los modelos de suavizamiento y ARIMA, la descomposici n estacional busca identificar los componentes estacionales de forma separada para mejorar la precisi n y permitir entender el comportamiento de la serie.

La descomposici n se construye a partir de un patr n (que estar  conformado por la tendencia, ciclo y variaci n estacional) y la suma o multiplicaci n de un error (correspondiente a la diferencia del valor pronosticado con el dato analizado).

En la figura posterior podemos ver que los modelos que contempla SPSS para la descomposici n, se dividen en:

- Multiplicativo. Al incrementarse la serie tambi n sucede lo mismo con el efecto de la variaci n. La f rmula general de la ecuaci n multiplicativa es:

$$Y_t = I_t \times T_t \times C_t \times e_t$$

- Aditivo. Es la descomposici n m s com n que podemos encontrar, su modelo matem tico es:

$$Y_t = I_t + T_t + C_t + e_t$$

En ambos modelos tenemos involucrados el componente estacional (I), la tendencia (T), un componente c clico (C) y el error aleatorio (e); todos ellos en para el per odo t. La combinaci n de cada elemento generar  un modelo distinto.

Espec ficamente el caso multiplicativo el error aleatorio puede ser incluso aditivo o no contar con el ciclo, de este modo podemos tener modelos multiplicativos tales como:

$$Y_t = I_t \times T_t \times C_t + e_t \qquad Y_t = I_t \times T_t + e_t \qquad Y_t = I_t \times T_t \times e_t$$

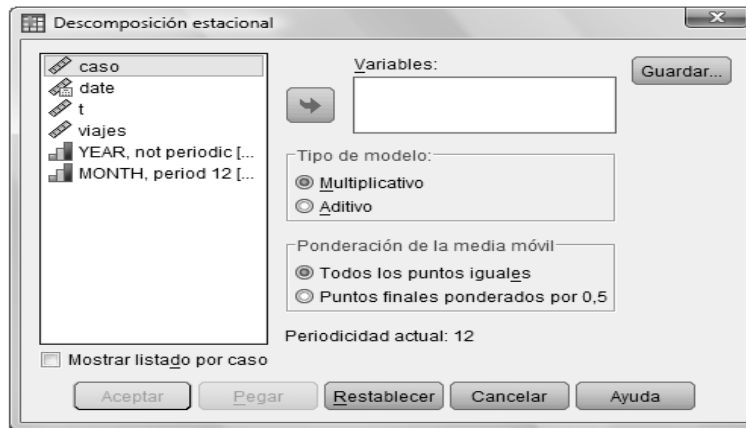


Fig. 54. Descomposición estacional

5.6 CREAR SERIE TEMPORAL

Los modelos que vimos en los temas anteriores los podemos obtener dentro del menú Analizar, sin embargo, una alternativa más que nos ofrece SPSS en cuestión de pronósticos es la opción “Crear serie temporal” del menú Transformar.

El propósito de esta opción es crear nuevas variables basadas en funciones de series temporales, el nombre que tendrá dicha variable estará compuesto por las primeras seis letras de la variable colocada en la casilla, seguida por un guión bajo.

Las funciones de serie temporal con las que cuenta SPSS son:

1. Diferencia (no estacional). El orden será el número de valores previos utilizados para calcular la diferencia. No olvidemos que el orden de la diferencia disminuirá en uno los datos de la serie de tiempo.
2. Diferencia estacional. El orden se basa en la periodicidad definida actualmente, para este tipo de diferencia se perderán de acuerdo al orden por la periodicidad definida.
3. Media móvil centrada. Se trata del promedio de un rango de valores de la serie; el valor de la amplitud estará dado por el número de valores utilizados para calcular el promedio.
4. Media móvil anterior. Se trata del promedio de un rango de las observaciones precedentes.
5. Medianas móviles. Tiene el mismo planteamiento que el método de medias móviles, el valor de la amplitud dependerá del número de datos de la serie.
6. Suma acumulada. Se trata de una sustitución de los datos por la suma acumulada de los valores precedentes, incluyendo el valor actual.
7. Retardo. Cada valor se sustituye por el que le preceda que estará determinado por el orden.
8. Adelanto. Es el caso inverso del retardo, de tal forma que hará sustitución con los valores posteriores.
9. Suavizado. Realiza un suavizado compuesto de medianas móviles de distintos tamaños y varias sucesiones.

De tal forma que dependerá de nosotros qué tipo de función usemos para crear la serie temporal.

5.7 EJEMPLO GENERAL

En esta ocasión, para ejemplificar lo visto en la unidad, utilizaremos los datos del Índice Nacional de Precios al Consumidor (INPC) de la sección de estadísticas del portal del Banco de México.

Dentro del apartado “Inflación”, tendremos dos subsecciones, daremos clic en “Inflación al Consumidor y UDIS”, después dentro de las Estructuras de Información en Inflación daremos clic en “Inflación”, finalmente seleccionaremos “Inflación mensual” y exportaremos los datos al formato de salida de Excel.

Los datos serán mensuales, correspondientes a la variación porcentual a partir de enero de 1973 hasta septiembre de 2010.

Cuando exportemos nuestros datos a SPSS, los formatos de fecha cambiarán a números, sin embargo, no debemos olvidar que tenemos que definir fechas, luego de esto, SPSS creará variables adicionales a las que tengamos dependiendo del tipo de formato que hayamos elegido (ver Fig. 55).

Fecha	SP30577	YEAR_	MONTH_	DATE
26665	1.45000000	1973		1 JAN 1973
26696	.83000000	1973		2 FEB 1973
26724	.88000000	1973		3 MAR 1973
26755	1.58000000	1973		4 APR 1973

Fig. 55. Variables de fecha

El siguiente paso será el gráfico de secuencia, usaremos como etiqueta del tiempo a la variable DATE_ y en variables colocaremos a “SP30577”.

En el gráfico de la Fig. 56 podemos observar que las variaciones de la inflación en el periodo 1973 a 2010 son relativamente pocas, además de que el crecimiento se ha ido estabilizando a partir del año 2005, esto podríamos adjudicárselo a las crisis económicas que hubo en los años 1982, 1983, 1994, 1995, entre otros.

Empezaremos analizando si nuestra serie de tiempo es estacionaria, es decir, que su media y varianza sean constantes en el tiempo y la función de autocorrelación sea independiente del tiempo.

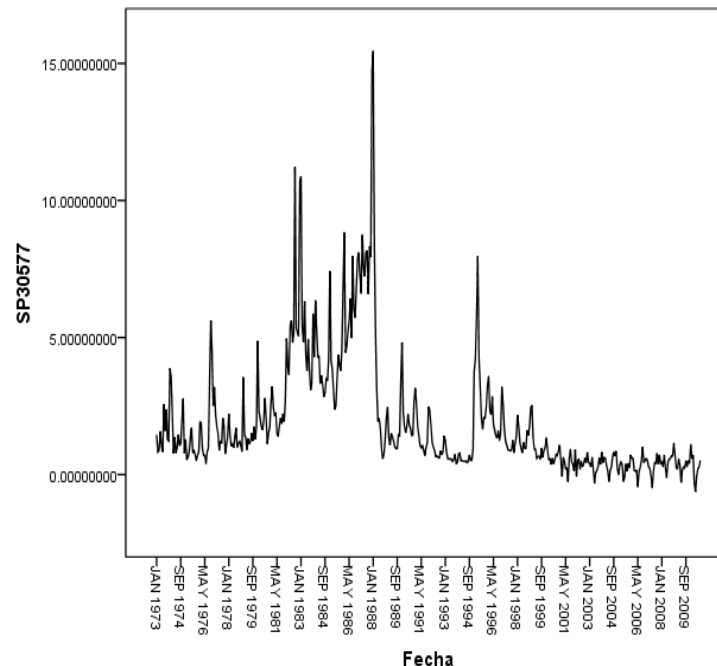


Fig. 56. Inflación 1973-2010

Examinando el gráfico de secuencia vemos que en efecto nuestra serie no es estacionaria, de modo que debemos aplicar alguna transformación para hacerla estacionaria, a esto podemos preguntarnos por qué la insistencia en que se cumpla este criterio, la respuesta es sencilla, ya que de esta forma la distribución de los datos será la misma para todos y por tanto facilita la elección para el modelo de pronóstico.

Ahora probaremos qué tan eficiente es el modelizador. Si bien no contamos con un modelo previo, sí tenemos las herramientas estadísticas para poder analizarlo, así que ¡empecemos!

A un lado de la opción del “Modelizador experto” tenemos el botón de Criterios, con el que configuraremos que se tomen en cuenta todos los posibles modelos, y que también se detecten automáticamente los valores atípicos.

Muy bien, para complementar el análisis del ajuste del modelo seleccionaremos el estadístico R cuadrado; en la pestaña de Guardar seleccionaremos “Valores pronosticados”.

La primer salida que nos arrojará SPSS será un cuadro de advertencia en el que nos explicará que se encontraron valores perdidos, por lo que el periodo de estimación se recortó (en nuestro caso el modelizador inició en diciembre de 1974)

La segunda sección corresponderá al resumen del modelo, como se muestra en la tabla siguiente:

Estadístico de ajuste	Media	ET	Mínimo	Máximo	Percentil							
					5	10	25	50	75	90	95	
R-cuadrado estacionaria	.891	.	.891	.891	.891	.891	.891	.891	.891	.891	.891	.891
R-cuadrado	.960	.	.960	.960	.960	.960	.960	.960	.960	.960	.960	.960
RMSE	.466	.	.466	.466	.466	.466	.466	.466	.466	.466	.466	.466
MAPE	39.994	.	39.994	39.994	39.994	39.994	39.994	39.994	39.994	39.994	39.994	39.994
MaxAPE	927.558	.	927.558	927.558	927.558	927.558	927.558	927.558	927.558	927.558	927.558	927.558
MAE	.319	.	.319	.319	.319	.319	.319	.319	.319	.319	.319	.319
MaxAE	2.485	.	2.485	2.485	2.485	2.485	2.485	2.485	2.485	2.485	2.485	2.485
BIC normalizado	-1.106	.	-1.106	-1.106	-1.106	-1.106	-1.106	-1.106	-1.106	-1.106	-1.106	-1.106

Tabla 29. Ajuste del modelo

Con el valor de RMSE, podemos identificar que el valor del ajuste es bueno ya que entre más pequeño sea mejor es el ajuste.

Asimismo, en la tabla de estadísticos ubicaremos el de r-cuadrado, que determina en qué porcentaje explica una variable a otra; en nuestros caso es el del 96%, que es bastante bueno.

Hasta el momento el modelo obtenido por SPSS ha sido bastante bueno, ya que los estadísticos de ajuste nos lo han revelado; como un complemento a esto obtendremos la gráfica de ambas series (ver Fig. 57).

De la gráfica posterior vemos que en verdad el modelo es bastante bueno, ya que con un 95% de confianza representa muy bien a la serie original.

Ahora podemos preguntarnos cuál es el modelo que obtuvo SPSS, ya que hasta ahora nos hemos centrado en qué tan bueno es el modelo, pero no sabemos si corresponde a una de las variaciones del modelo ARIMA, o si se trata de un tipo de suavizamiento. Es fácil saber este dato, ya que viene en la tabla que describe el modelo, el cual para este ejemplo fue un modelo ARIMA (0, 1, 2)(0, 1, 1).

Modelo	Número de predictores	Estadísticos de ajuste del modelo	Ljung-Box Q(18)			Número de valores atípicos
		R-cuadrado estacionaria	Estadísticos	GL	Sig.	
SP30577-Modelo_1	1	.891	24.666	15	.055	26

Tabla 30. Estadísticos del modelo

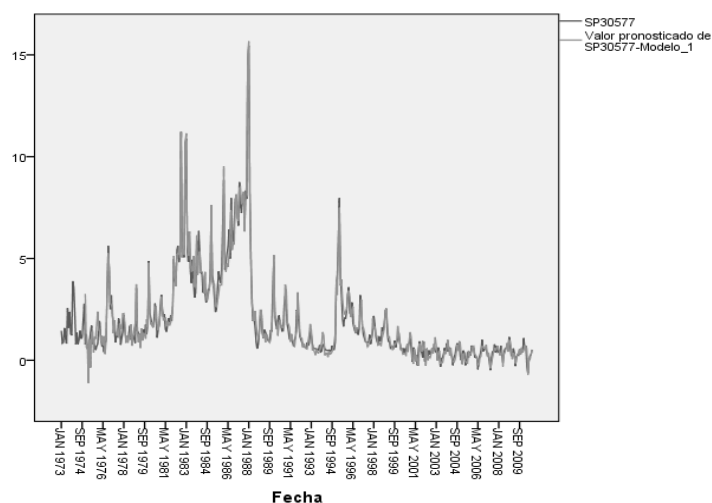


Fig. 57. Serie vs. modelizador

Con esto si deseamos implementar este modelo en otro software bastará con configurar las opciones necesarias para obtenerlo.

5.8 EJERCICIO COMPLEMENTARIO

Ingresemos de nuevo al portal de Gapminder, tal como lo hicimos en el ejemplo general del capítulo 4. Una vez ahí, busquemos el indicador “People living with HIV”, y daremos clic en el icono del archivo de Excel para que podamos guardar los datos en nuestro equipo.

Dejaremos en una nueva pestaña los datos de Ethiopia, con dos columnas una para el año iniciando en 1979 hasta el 2007; y la segunda columna corresponderá al número de personas con SIDA en el país de Etiopía.

En este ejemplo, usaremos la opción “Crear serie temporal” para convertir nuestras variables a aquellas que tengan la distribución dada por la función en el listado que ya mencionamos en el tema 5.6.

Usaremos la función de suavizado y colocaremos la variable “Ethiopia” en la casilla de variable y luego cambiaremos la función que nos da por default por la de Suavizamiento, por último pulsaremos el botón “Cambiar” e inmediatamente se generará una nueva variable, y en nuestro visor de resultados obtendremos una tabla de nombre “Serie creada”, la cual contiene las especificaciones de la nueva variable, tales como: nombre, casos no perdidos y válidos y la función con que fueron creados.

¿Qué podemos decir de esto? Tenemos una nueva variable a la que se le aplicó una función, pero ¿qué relación tiene con nuestros datos originales? Obtengamos la gráfica de ambas series, tanto la original como la resultante de la función de Suavizado para observar qué comportamiento tienen una con otra.

Aproximadamente son iguales los valores del suavizamiento con los valores originales de la serie, esto es porque las funciones de serie de tiempo tienen la finalidad de representar observaciones para un momento diferente con una duración del tiempo uniforme.

Finalmente qué podemos concluir observando la gráfica de la serie temporal (ver Fig. 58), recordemos que nuestros datos representan el número de personas que existen con SIDA en Etiopía que van desde el año 1979 hasta el 2007.

A primera vista pareciera ser una gráfica que tiene el comportamiento de una función exponencial, sin embargo, la forma en que ha crecido se asemeja más a la curva S, la cual comienza creciendo lento (tal como vemos del año 1979 al año 1989), después rápido (de los años 1991 a 1999) y después su crecimiento es de nuevo lento (correspondiente a los años 2001 a 2007).

Lo anterior nos indica que en sus inicios el SIDA, en Etiopía, no tenía la magnitud que el periodo 2001-2007, sin embargo, ¿a qué se debe este fenómeno? La respuesta, en general para todo país, es que depende de la forma de vida, así como de la información con la que cuenta acerca de lo que es el SIDA, además, resulta alentador ver que para el futuro el número de personas infectadas ya no se disparará, a menos de que se siga este comportamiento de la gráfica S, es decir, que tenga de nuevo un crecimiento lento, luego un crecimiento exponencial, seguido de un fase de estabilización.

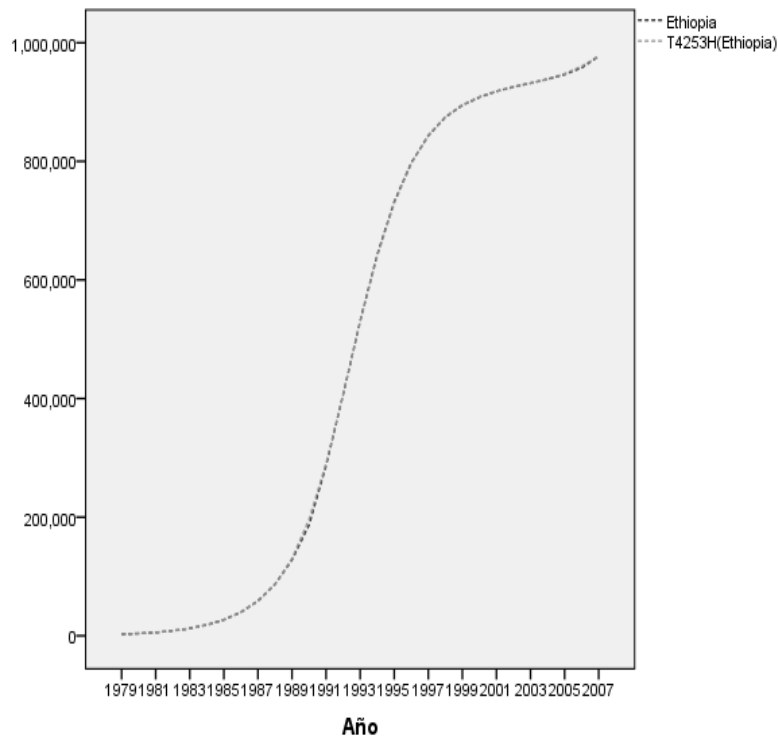


Fig. 58. Gráfico vs.suavizamiento

CONCLUSIONES

Todo lo que empieza tiene que terminar en algún momento, así que ahora después de que vimos fórmulas, gráficos, tablas, cuadros de diálogos, etc. ¿qué es lo que nos falta? Sólo el hecho de concluir sí, pero... ¿de todo esto qué podemos concluir?

Al inicio del presente documento se estableció el objetivo: Facilitar el aprendizaje e interpretación de la estadística para motivar el estudio de la misma, con el apoyo del software estadístico SPSS. Esto con la premisa de que si la estadística para los alumnos resulta muy teórica y tediosa debido a la realización de análisis como obtener promedios, encontrar valores en tablas, regresiones lineales; entonces el uso de un software estadístico para aplicaciones prácticas reales facilitará el aprendizaje e interpretación de la estadística.

¿Qué nos encontramos en el camino? Una serie de temas que con apoyo de SPSS obtuvimos cálculos como la media, desviación, pruebas de hipótesis, de una forma más rápida. Dichos temas estuvieron enfocadas principalmente a manipular las funciones del software para después centrarnos en lo que corresponde a la obtención, interpretación y análisis de los resultados obtenidos. Lo anterior con la finalidad de hacer una conclusión adecuada o de incluso inferir las posibles causas; dicha conclusión tratamos de que fuera entendible para cualquier persona, sin la necesidad de remitirse a los cálculos con los que fue obtenido.

En la justificación vimos el por qué realizar este trabajo, tomando en cuenta que la estadística es una disciplina usada en diversas licenciaturas, y que el problema de su aprendizaje se ha encontrado en varios países, a lo que Behar [5] comentó que es posible abordar esta problemática enseñándole al alumno que puede resolver algún problema que sea de su interés.

En el capítulo 1 nos introducimos al manejo del software, mientras que en el capítulo 2 empezamos con la estadística descriptiva así como la forma adecuada de dar una conclusión después de analizar los datos tomados de la realidad, y todo esto con la ayuda de las herramientas que nos brinda SPSS.

En el capítulo 3, conocimos los elementos y tipos de pruebas de hipótesis, así como las herramientas con las que cuenta SPSS en pruebas paramétricas y no paramétricas, como ejemplo, vimos que si bien se han hecho investigaciones y dado conclusiones a partir del uso de pruebas de hipótesis, también es posible complementar esa visión con elementos básicos de éstas pruebas para mejorar el estudio en cuestión.

Para el capítulo 4, vimos los temas de la correlación y regresión; los cuales fueron estadísticos fáciles de obtener, y generar así los modelos, ya que bastó con unas cuantas configuraciones en SPSS para definir cómo fue la correlación, o en su caso, si el modelo de regresión lineal es el más adecuado en cuanto al ajuste de éste con los datos ya observados en la serie original.

En el capítulo 5 vimos los modelos con los que cuenta SPSS para hacer pronósticos de series de tiempo, dependiendo si esta era estacionaria, o no. Además, revisamos qué ajuste tuvo el modelo contenido con el modelizador experto, y cómo ha estado de 1979 a 2007 el SIDA en Etiopía.

Cabe resaltar que no sólo fue cuestión de describir el uso del software y poner un ejemplo, se trató de que una vez visto lo correspondiente al capítulo se seleccionara un tema para hacer las pruebas estadísticas correspondientes al capítulo, además de la interpretación de los resultados; cada ejemplo fue completamente auténtico, de tal forma que fuera como un trabajo en campo en el que tocara revisar ciertos aspectos, acercándose lo más posible al trabajo que se hace en la investigación, en las empresas e incluso a nivel licenciatura.

Además de lo anterior, también se realizó un pequeño experimento en el que se contrastó la efectividad que tiene usar SPSS para la resolución del ejercicio complementario del capítulo 2. De dicha prueba se constató que no existe diferencia significativa entre usar un método u otro, sin embargo, con la aplicación de una encuesta se pudo observar que, en efecto, los ejemplos tomados de la realidad y el manejo de software ayudan al aprendizaje de la estadística. Asimismo, se identificó que no sólo son éstos factores los

que determinarán el aprendizaje, sino que en conjunto con las clases presenciales, la práctica, utilidad que le den, y las bases que tenga el usuario le permitirán aprender más fácilmente la estadística.

Con lo expuesto anteriormente vemos que fue posible facilitar el aprendizaje e interpretación de la estadística, si bien no en un cien por ciento, éste material sí puede usarse como una herramienta adicional a las clases convencionales, de modo que se cree un aliciente para su estudio, es decir, quitando las barreras hacia el temor de las matemáticas; y todo esto bajo la hipótesis de que un software estadístico sea nuestra herramienta y el uso de aplicaciones prácticas reales es una de las posibles soluciones a un problema que ha ocurrido en diversos países.

¿Cuáles fueron los inconvenientes que nos encontramos en el camino? Lo más recurrente fue el hecho de encontrar un ejemplo suficientemente manejable de acuerdo al capítulo referido, ya que este debió de ser simple de entender, interesante y al cual se le puedan aplicar las pruebas expuestas en el capítulo correspondiente; es por ello que cada uno fue completamente distinto. Si bien se llegó a utilizar información que no seguía la misma línea del resto de los ejemplos fue con la finalidad de evitar la monotonía que suelen tener algunos libros al plantear situaciones no reales, además de no llevarse como un simple manual de usuario.

Por lo que se refiere a la prueba realizada, el inconveniente fue no haber contado con una separación física de los grupos, no haber tenido más de una sesión, esto debido a que no se contó con una buena planeación.

¿Qué queda por hacer? Lógicamente este trabajo es sólo un pequeño paso a nombre de todo aquel estudiante que se haya, literalmente, “dado de topes” por tener que llevar una materia relacionada con la estadística, por toda aquella persona que en su vida profesional se sienta limitada al hacer un análisis estadístico, por aquellas personas a las que nos resulta más interesante manipular un software que nos sea de herramienta para la resolución de problemas reales, en fin, queda claro que si bien hay diversos materiales no es sólo cuestión de hacerlos, sino también de saberlos usar, y de difundirlos.

En concreto, el siguiente paso conlleva a contar con un manual-libro en la red, que tenga el enfoque del aprendizaje cualitativo que involucre y motive al estudiante a investigar y querer adentrarse más; un curso lo suficientemente original que pueda estar a la mano de cualquier estudiante interesado en aprender estadística.

Lo anterior está comenzando a ser una realidad, dentro del proyecto PAPIME PE 300309 se dio el alojamiento para el curso PASW, el cual aún está en construcción, y pretende además usar el contenido del presente material, tener como complemento: funciones no vistas, páginas de interés, chats, foros, artículos relacionados con la estadística.

Por lo que ahora no queda más que terminar de elaborar este curso, y poder implementarlo, sin olvidar que el objetivo es: “Facilitar el aprendizaje e interpretación de la estadística para motivar el estudio de la misma, con el apoyo del software estadístico, SPSS”.

FUENTES DE INFORMACIÓN


- [1] Banxico. (s.f.). *Banco de México*. Recuperado el 2010, de <http://www.banxico.org.mx/>
- [2] Batanero, C. (s.f.). Los retos de la cultura estadística. *Universidad de Granada, España*.
- [3] Batanero, C., & Díaz, C. (2010). El papel de los Proyectos en la Enseñanza y Aprendizaje de la Estadística. *Universidad de Granada, España*, 22.
- [4] Behar, R., & Ojeda, M. (s.f.). El problema de la Educación Estadística: Perspectiva desde el Aprendizaje. *Ingeniería y Competitividad*, 7.
- [5] Behar, R., & Ojeda, M. (s.f.). La problemática y el aprendizaje de la estadística en la educación superior. 11.
- [6] Behar, R., Grimas, P., & Ojeda, M. M. (2005). Educación Estadística Ámbito Universitario: Algunas Reflexiones. 29.
- [7] Bowerman, Bruce, et. al. (2007). Pronósticos, series de tiempo y regresión. México: Cengage Learning.
- [8] Camacho, R. (2006). *Estadística con SPSS para Windows versión 12*. México: AlfaOmega Ra-Ma.
- [9] Canavos, C. (1998). *Probabilidad y Estadística. Aplicaciones y métodos*. México: Mc Graw-Hill.
- [10] Cantú, P., & Gómez, L. (s.f.). El valor de la estadística en la salud pública. *Centro de Ginecología y Obstetricia de Monterrey, S. A. de C. V.*, 4.
- [11] Company, I. (s.f.). *SPSS*. Recuperado el 2010, de <http://www.spss.com/>.
- [12] Galavosky, L. (2004). *Del aprendizaje significativo al aprendizaje sustentable. Parte 1: El modelo teórico*. Buenos Aires.
- [13] Gallese, E. (2000). Problemática sobre la enseñanza y aprendizaje de la estadística en carrera no estadísticas.
- [14] Gapminder. (s.f.). *Gapminder*. Recuperado el 2010, de <http://www.gapminder.org/about-gapminder/our-mission/>
- [15] García, M. (2007). Comunicación y aprendizaje electrónico: la interacción didáctica en los nuevos espacios virtuales de aprendizaje.
- [16] González Videgaray, M. d. (2009). *Pronósticos: Metodología de Box-Jenkins*. México.
- [17] González Videgaray, M. d. (2008). *Series de tiempo II*.
- [18] Hernández, J. D. (1989). Aplicación de la estadística a la investigación administrativa. *Acta mexicana de ciencia y tecnología* , 8.
- [19] Incorporation, S. (2007). *Manual del usuario de PASW Statistics Base 17*.
- [20] Incorporation, S. *SPSS 10. Guía para el análisis de datos*.

- [21] INEGI. (2007). *BIE. Banco de Información Estadística*. Obtenido de <http://dgcnesyp.inegi.gob.mx/>
- [22] Levesque, R. (2007). *Programming and Data Management for PASW Statistics 18. A guide for PASW Statistics and SAS Users*. IBM Company.
- [23] MAC, J. d. (2005). *Plan de estudios de la licenciatura en Matemáticas Aplicadas y Computación*. México.
- [24] Pérez Guerrero, Y. (s.f.). De la Barreda Bautista B. Comunidades Virtuales de Aprendizaje como herramienta didáctica para el apoyo de la labor docente.
- [25] Programa, J. d. (2004). *Programa de estudios de las asignaturas*. México.
- [26] Spiegel, M. (1976). *Teoría y 760 problemas resueltos*. México: McGraw-Hill.
- [27] Spiegel, M. (1992). *Teoría y problemas de probabilidad y estadística*. México: Mc Graw-Hill.
- [28] Tecnológico, C. d. (2005). *Manual SPSS*.
- [29] Ulloa, V. (2006). *Estadística aplicada a la comunicación. Antología*. México: FES acatlán.
- [30] UNAM. (2008). *Portal de Estadísticas Universitarias*. Recuperado el 2010, de <http://www.estadistica.unam.mx/>
- [31] Walpole, M. (1999). *Probabilidad y Estadística para ingenieros*. México: Prentice-Hall Hispanoamericana.

ANEXO A. SINTAXIS DE SPSS

Como ya habíamos mencionado, SPSS cuenta con una herramienta con la que se puede editar y generar los resultados de las pruebas estadísticas mediante sintaxis, es decir, instrucciones nativas del software.

¿Cuál es el objeto de contar con sintaxis y comandos si ya tenemos los cuadros de diálogo? Pues bien, hay varias instrucciones que son exclusivas de la sintaxis, además, podemos ejecutar más de un comando a la vez, lo que es posible aprovechar una sola corrida para varios procesos, que es lo que en el mundo de los negocios minimiza los costos; pues bien empezemos a conocer esta otra herramienta que nos brinda SPSS.

En la Fig. 59 podemos ver el botón  que es con el que se ejecutan las instrucciones o bien podemos usar la combinación de teclas Ctrl+R, para ambos casos será necesario que seleccionemos el código a ejecutar.

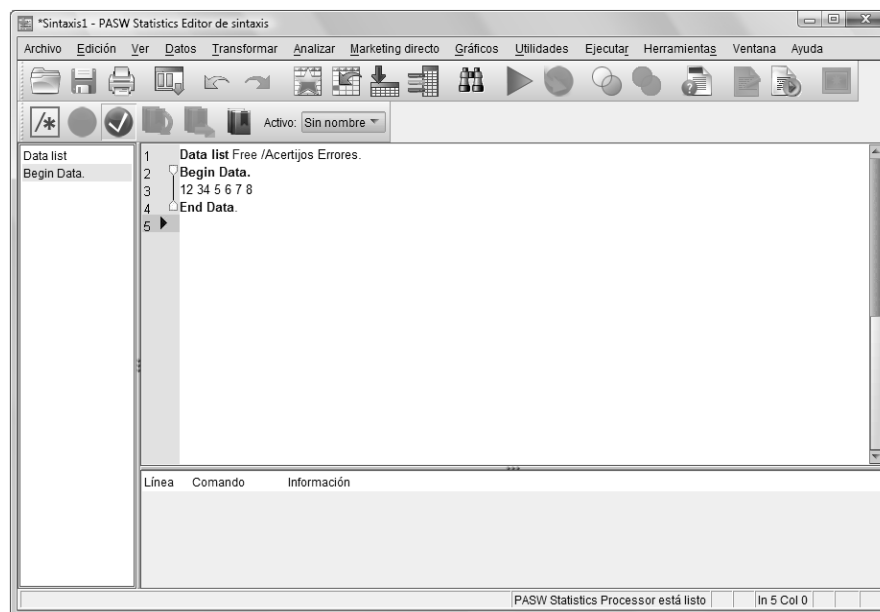


Fig. 59. El editor de sintaxis

Con respecto a las palabras reservadas, éstas pueden estar de color azul, vino, o verde.

Con el código anterior creamos dos variables, para las cuales sus valores estarán dados por los comandos Begin y End Data.

Una vez que hayamos corrido nuestra sintaxis se abre una visor de resultados en el que de haber errores nos los especifica, y sino únicamente se transcribe dicho código y se obtiene el resultado, dependiendo de si es una tabla o un gráfico.

Debido a que generalmente tendremos nuestros datos capturados, nos centraremos en la manipulación de ellos, así que, ¡continuemos!

Para guardar nuestro archivo de datos usaremos el comando:

```
SAVE OUTFILE='miarchivo.sav'.
```


En este caso se guardó automáticamente en la carpeta de SPSS; sin embargo, podemos escribir toda la ruta para que este se guarde en el lugar que deseemos. Para el caso opuesto en que queramos trabajar en algún archivo que tengamos previamente nos ayudaremos del comando:

```
GET FILE='miarchivo.sav'.
```

Si se tratara de un archivo en Excel, tendremos:

```
GET DATA
/TYPE=XLS
/FILE='C:\Users\Alpha\Documents\T\U1\egresados.xls'
/SHEET=name 'Hoja1'
/CELLRANGE=full
/READNAMES=on
/ASSUMEDSTRWIDTH=32767.
```

Con este conjunto de comandos abriremos el archivo “egresados.xls”, específicamente la hoja1 de la que se tomarán todos los datos tomándose en cuenta que los datos de ese archivo cuentan con nombres definidos y por último se define el ancho de columna.

Del código anterior, podemos destacar también, que a pesar de ser varios comandos estos se dividen por “/”, y al término de una instrucción será necesario que coloquemos un “.”, ya que será el terminador que define un conjunto de instrucciones.

Sigamos, para la creación de variables tenemos el comando COMPUTE, analicemos qué obtendremos de la corrida del código siguiente:

```
COMPUTE Egre_matri=Egresados/Matrícula.
VARIABLE LABELS Egre_matri 'Egresados vs ingreso'.
RECODE Período (1969 thru Highest=1) (1991 thru Highest=2) (2001 thru Highest=3).
```

En primer punto creamos a la variable “Egre_matri” a partir de la división de otras dos variables con la iiqueta de variable “Egresados vs ingreso”. La segunda parte corresponde a la recodificación de la variable Período de acuerdo a las condiciones dadas en cada uno de los paréntesis.

En la tabla 31, tenemos los comandos que corresponden al tipo de variable.

NOMINALES	ORDINALES	DE INTERVALO
FREQ región /BARChart /STAT=MODE.	FREQ estatus /PIEChart /STAT=MEDIANA, MAXIUM, MINIUM, RANGE.	FREQ peso, ci, estatura /HISTOGRAM=NORMAL /STAT=DEFAULT, KURTOSIS, RANGE.

Tabla 31. Comandos por tipo de variable

Podríamos continuar analizando comando por comando, pero ahora que ya estamos picados nos tocará aprenderlos poco a poco, para lo cual nos ayudará SPSS, ya que cada vez que hagamos algo a partir de los cuadros de diálogo se escribe automáticamente el código de sintaxis en el visor de resultados, y para el caso en que queramos saber más acerca de las bondades de trabajar con el lenguaje nativo de SPSS contamos con cursos que ofrece la propia empresa IBM o revisar alguna de las guías para programación en el modo de sintaxis, las cuales están integradas en el menú de ayuda de SPSS.

ANEXO B. ¿QUÉ HAY DE NUEVO?... SPSS 19

SPSS IBM se ha preocupado por estar a la vanguardia y ampliar las características del software en cuestión, de modo que ya cuenta con la nueva versión de SPSS, para la cual destacan las siguientes funcionalidades que ha integrado en este SPSS 19:

- Desempeño más rápido – Desempeño 200% más rápido al crear reportes con grandes bases de datos o varias tablas pequeñas.
- Modelos Lineales Automáticos – Una automática y sencilla forma de construir poderosos modelos lineales.
- Editor de Sintaxis – Más de una docena de mejoras que hacen más sencillas la escritura de sintaxis.
- Ponderación de datos - Fácilmente pondere nuevos datos de clientes, acceda a modelos pre construidos y tenga una interfaz directa con datos de Sales forcé.com.
- Mostrar/ocultar filas, columnas o etiquetas de forma selectiva para resaltar hallazgos importantes
- Ayuda de tareas con instrucciones paso a paso.
- Crear y guardar las especificaciones del análisis para utilizarlas en tareas repetitivas o en procesos desatendidos (sin supervisión de un usuario).
- Controlar la división de las tablas con funciones de paginación e impresión.
- Consultar las explicaciones de los términos estadísticos contenidas en el glosario estadístico en pantalla.
- El comando HOST permite aprovechar la funcionalidad del sistema operativo en Statistics. Este comando permite a las aplicaciones “escapar” del sistema operativo y ejecutar otro programas de forma sincronizado con la sesión de Statistics.
- Verificación de reglas en gráficos secundarios SPC.

Además, no olvidemos que también cuenta con las características que fueron explicadas a lo largo del documento, así que sólo resta esperar para ver con qué nos sorprenderán en la próxima actualización.

ÍNDICE DE FIGURAS

Fig. 1.	Asistente para introducir datos	- 13 -
Fig. 2.	Editor de datos	- 14 -
Fig. 3.	Visor de resultados	- 14 -
Fig. 4.	Etiqueta de valor	- 16 -
Fig. 5.	Abrir datos	- 18 -
Fig. 6.	Calcular variable	- 19 -
Fig. 7.	Recodificar en distintas variables	- 20 -
Fig. 8.	Valores antiguos a nuevos	- 20 -
Fig. 9.	Recodificación automática	- 21 -
Fig. 10.	Agrupación visual	- 22 -
Fig. 11.	Segmentar archivo	- 23 -
Fig. 12.	Seleccionar casos	- 24 -
Fig. 13.	Ponderar casos	- 25 -
Fig. 14.	Guardar datos como	- 26 -
Fig. 15.	Recodificación	- 27 -
Fig. 16.	Frecuencias	- 30 -
Fig. 17.	Estadísticos	- 30 -
Fig. 18.	Coefficientes	- 32 -
Fig. 19.	Seguridad Microsoft Excel	- 35 -
Fig. 20.	Editor de gráficos	- 36 -
Fig. 21.	Propiedades del gráfico	- 36 -
Fig. 22.	Exportando...	- 37 -
Fig. 23.	Histograma Servicio Local	- 38 -
Fig. 24.	Diagrama de caja clientes celular	- 41 -
Fig. 25.	Variables	- 42 -
Fig. 26.	Nivel vs. Género	- 43 -
Fig. 27.	Sistema vs. Género	- 44 -
Fig. 28.	Porcentajes con respecto a la media y desviación estándar con una población normalmente distribuida. (Fuente: Bowerman, 2007)	- 44 -
Fig. 29.	Medias	- 49 -
Fig. 30.	Prueba T para una muestra	- 51 -
Fig. 31.	Opciones	- 51 -
Fig. 32.	Prueba T para muestras independientes	- 53 -
Fig. 33.	Prueba T para muestras relacionadas	- 55 -
Fig. 34.	ANOVA de un factor	- 56 -

Fig. 35.	Prueba de chi-cuadrado	- 59 -
Fig. 36.	Prueba binomial	- 60 -
Fig. 37.	Prueba de rachas	- 61 -
Fig. 38.	Prueba K-S para una muestra	- 62 -
Fig. 39.	Pruebas para 2 muestras independientes	- 65 -
Fig. 40.	Pruebas para varias muestras independientes	- 66 -
Fig. 41.	Pruebas para dos muestras relacionadas	- 68 -
Fig. 42.	Pruebas para varias muestras relacionadas	- 70 -
Fig. 43.	Tipos de correlación	- 79 -
Fig. 44.	Correlaciones bivariadas	- 80 -
Fig. 45.	Regresión lineal	- 83 -
Fig. 46.	Definir fechas	- 92 -
Fig. 47.	Ejemplos series de tiempo (FUENTE: Bowerman, 2007)	- 93 -
Fig. 48.	Gráficos de secuencia	- 93 -
Fig. 49.	Serie Viajes	- 94 -
Fig. 50.	Autocorrelaciones	- 95 -
Fig. 51.	Periodograma	- 95 -
Fig. 52.	Modelizador de series temporales	- 98 -
Fig. 53.	Criterios de ARIMA	- 99 -
Fig. 54.	Descomposición estacional	- 102 -
Fig. 55.	Variables de fecha	- 103 -
Fig. 56.	Inflación 1973-2010	- 103 -
Fig. 57.	Serie vs. modelizador	- 105 -
Fig. 58.	Gráfico vs.suavizamiento	- 106 -
Fig. 59.	El editor de sintaxis	- 111 -

ÍNDICE DE TABLAS

Tabla 1.	Atributos	- 27 -
Tabla 2.	Acordeón	- 34 -
Tabla 3.	Estadísticos 1	- 37 -
Tabla 4.	Estadísticos 2	- 38 -
Tabla 5.	Frecuencias con segmentación	- 42 -
Tabla 6.	Error tipo I y II	- 48 -
Tabla 7.	K muestras independientes de k poblaciones	- 56 -
Tabla 8.	Variables	- 71 -
Tabla 9.	Informe	- 71 -
Tabla 10.	Prueba para una muestra	- 72 -
Tabla 11.	Prueba T de muestras independientes	- 73 -
Tabla 12.	Prueba T para muestras relacionadas	- 73 -
Tabla 13.	Prueba de rachas	- 74 -
Tabla 14.	Prueba K-S para una muestra	- 74 -
Tabla 15.	Prueba de Wilcoxon	- 75 -
Tabla 16.	K-S de los grupos	- 77 -
Tabla 17.	Prueba T: promedio calificaciones y tiempo invertido	- 77 -
Tabla 18.	Prueba de Moses	- 78 -
Tabla 19.	Correlación de Pearson	- 86 -
Tabla 20.	Resumen del modelo	- 87 -
Tabla 21.	ANOVA	- 87 -
Tabla 22.	Coefficientes	- 87 -
Tabla 23.	Correlaciones parciales	- 89 -
Tabla 24.	Resumen del modelo múltiple	- 89 -
Tabla 25.	Coefficientes múltiple	- 89 -
Tabla 26.	Peso, estatura, edad	- 89 -
Tabla 27.	Resumen modelos	- 90 -
Tabla 28.	ANOVA modelos	- 90 -
Tabla 29.	Ajuste del modelo	- 104 -
Tabla 30.	Estadísticos del modelo	- 105 -
Tabla 31.	Comandos por tipo de variable	- 112 -