UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE ESTUDIOS SUPERIORES ZARAGOZA
DIVISIÓN DE CIENCIAS QUÍMICO BIOLÓGICAS

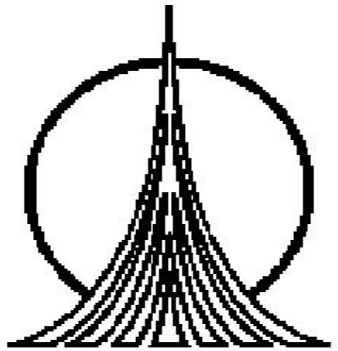
ACADEMIA DE ESTADÍSTICA Y CÓMPUTO CIENTÍFICO
Serie: Comunicaciones en Estadística y Cómputo Científico

Estadística Práctica para el Análisis de Datos

SPSS. Una herramienta para el análisis estadístico de datos

Armando Cervantes Sandoval
Patricia Rivera García
Juana María De la Paz López

2004



Estadística Práctica para el Análisis de Datos

SPSS. Una herramienta para el análisis estadístico de datos

Facultad de Estudios Superiores ZARAGOZA ISBN: 970-32-1260-3

PAPIME: EN216403

Presentación

Este material es una guía práctica para utilizar el software de análisis estadístico SPSS (*Statistical Package Social Science),* como una herramienta para el análisis de datos. Con ese fin se hace una revisión del "paquete" en un formato de curso, donde se describe paso a paso como utilizar y aplicar algunas de las muchas opciones básicas disponibles.

Enfocado al uso del software, sin perder formalidad y con el rigor necesario, en cada capítulo se presenta una muy breve explicación estadística, se indica como realizar el análisis correspondiente en SPSS y mediante ejemplos se dan algunos criterios que apoyan la interpretación de resultados.

Está dirigido a, quienes sabiendo que técnica estadística aplicar requieren de una herramienta confiable que se haga cargo del trabajo de cálculo numérico. No se pretende que sea un libro de estadística, tampoco un recetario de estadística y muchos menos un manual de SPSS; simplemente se busca que sea una herramienta de apoyo en el quehacer académico de estudiantes, profesores e investigadores. Que los motive e invite a revisar material más formal.

Para quienes no sepan o no tengan claro que técnica estadística aplicar pueden consultar alguno de los siguientes materiales:
- Cervantes S. Armando., 1998, **Métodos estadísticos, una herramienta más en la metodología de investigación**, Tópicos de Investigación y Posgrado, V(4):225-232.
- Mendoza Nuñez, V. M. y Sánchez Rodríguez, M. A., 2001, **Análisis y difusión de resultados científicos**, 1ª. Edición, FES Zaragoza, UNAM.
- Marques Dos Santos, M. J., 2002, **Guía para el uso de métodos estadísticos**, D'vεrit@σ, 1(3):2-5.

Se recomienda utilizar éste material directamente en el SPSS, rehaciendo los ejemplos y ejercicios.

Se agradece la revisión y correcciones realizadas por: M. en C. María José Marques Dos Santos, Biól. María del Carmen Galindo de Santiago (en especial por el capítulo de regresión), M. en C. José Vicente Rosas Barrientos y Mtro. Luis A. Mora Guevara, cuyas aportaciones verdaderamente ayudaron a mejorar la calidad de éste material.

Un reconocimiento especial a los alumnos del curso que dieron origen a las primeras versiones de estas notas (con todo respeto y afecto: alumnillos de indias), así como a los de cursos posteriores, por su paciencia y cuidado para detectar y ayudar a realizar algunas correcciones.

Finalmente, asumimos la responsabilidad de cualquier error, omisión o mala interpretación que, totalmente sin querer, se presente. Agradeciendo todas las correcciones, comentarios y sugerencias al correo electrónico: arpacer@servidor.unam.mx.

Armando Cervantes Sandoval

Patricia Rivera García

Juana María De la Paz López

ÍNDICE

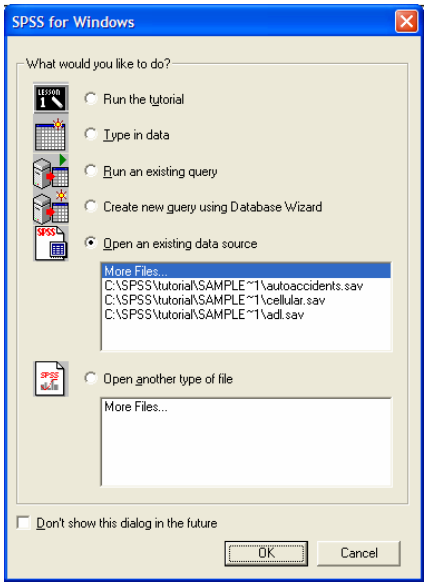
Capítulo 1

Conociendo el entorno de trabajo SPSS versión 11

Para iniciar una sesión de trabajo primero hay que asegurarse de que SPSS esté instalado en su computadora.

El primer paso consiste en "abrir el paquete", lo cual se puede hacer de dos formas:

1. Seguir la secuencia: **INICIO -> PROGRAMAS -> SPSS for Windows**
2. En algunos casos se puede tener un acceso directo, por lo que sólo basta con dar un **doble clic en el icono o botón correspondiente** y listo.



Al entrar a SPSS puede aparecer la siguiente caja de diálogo, la cual recomendamos **cancelar** y **activar la opción NO MOSTRAR ESTE DIALOGO EN EL FUTURO** (Don't show this dialog in the future)**.**

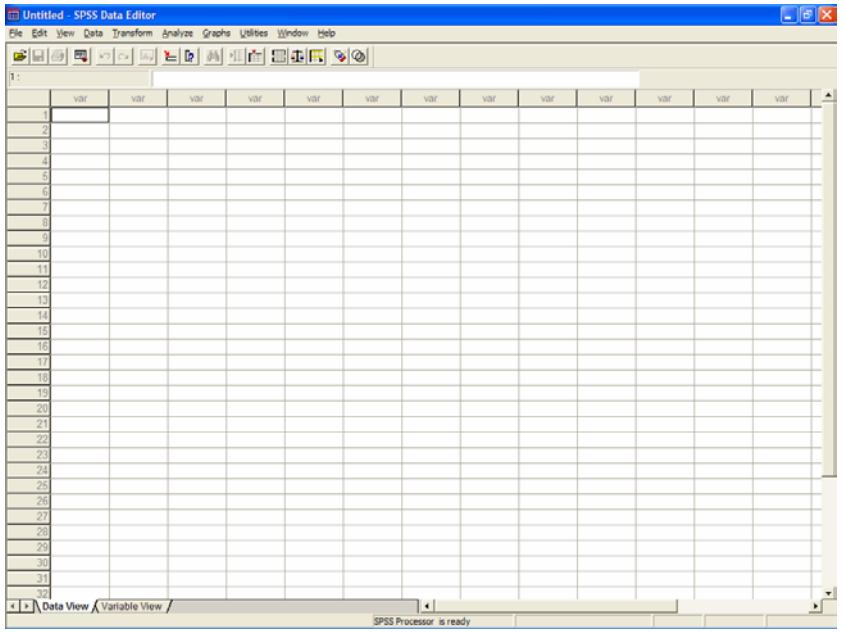
Una vez dentro del "paquete" lo primero es saber como salir. Esto se hace con:

1. Seleccionar de la barra de menú la opción **FILE** y luego **EXIT**.
2. En la barra de control dar un clic sobre el **icono de cerrar**, generalmente identificado por un feo tache (X). Usualmente hasta arriba y a la derecha.
3. En la misma barra de control, **dar un doble clic sobre el icono que identifica al paquete**, o un **clic y cerrar** (hasta arriba y a la izquierda).

Ahora si, listo para "ingresar" datos y empezar a trabajar. Lo primero que aparece es la ventana para captura y edición de datos, cuyo aspecto es muy semejante al de una hoja de cálculo, ubicando cada con un número de fila y un descriptor de columna.

Cada **celda** del editor puede almacenar **texto** o **valores numéricos**, pero a diferencia de una hoja de cálculo no puede almacenar fórmulas.

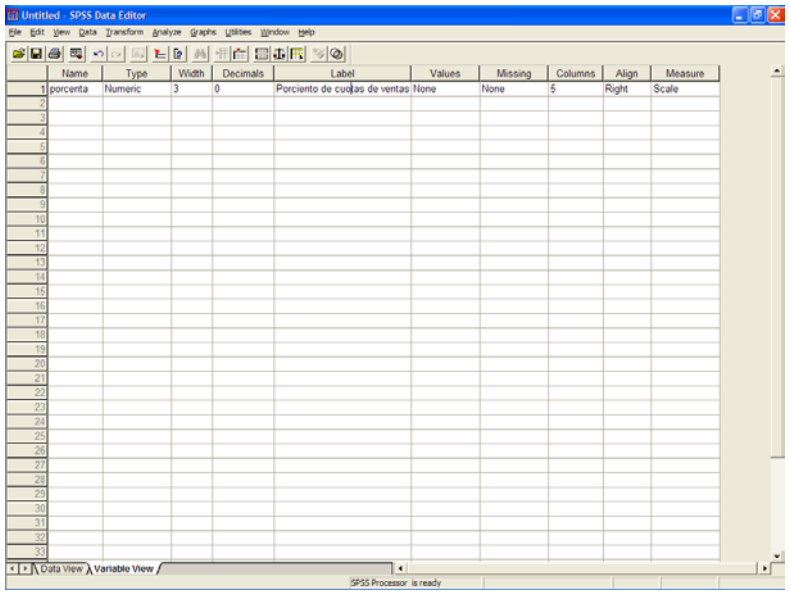
Este formato permite capturar los datos en una hoja de cálculo como Excel o en cualquier software que maneje un formato de celdas: STATGRAPHICS, MINITAB o algunos módulos de SAS, todos estos, paquetes estadísticos.



Para personalizar los datos se debe trabajar con las opciones tipo fólder localizadas en la parte inferior izquierda del editor de datos: **Data View**, la opción por omisión y **Variable View** la opción para darle nombre a las variables, definir tipo de datos y algunas otras características.

En la vista de variables (Variable view), se le puede dar nombre a todas y cada una de las variables en el conjunto de datos; definir el tipo de datos, generalmente de tipo numérico o texto (string), también permite definir el

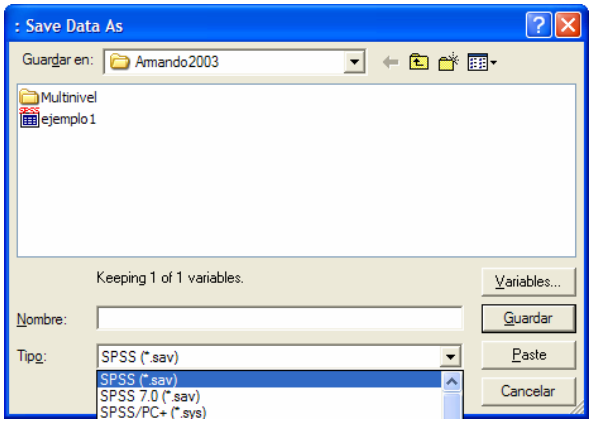
tamaño y número de decimales; una etiqueta para la variable; etiquetas para los diferentes valores de una variable (por ejemplo: 1 = femenino y 2 = masculino); definir valores faltantes; tamaño de la celda en número de columnas; la alineación de los datos; finalmente la escala de medición (nominal, ordinal o scale). Para realizar modificaciones hay que dar un clic en la opción correspondiente, apareciendo posibilidades de selección que permiten definir las condiciones adecuadas de trabajo.



Ejemplo 1. En un cierto mes, quince vendedores alcanzaron: 107, 90, 80, 92, 86, 109, 102, 92, 353, 78, 74, 102, 106, 95 y 91 por ciento de sus cuotas de venta.
(Estadística Elemental, John E. Freund y Gary A. Simon, 1992, Prentice Hall, pág. 56).

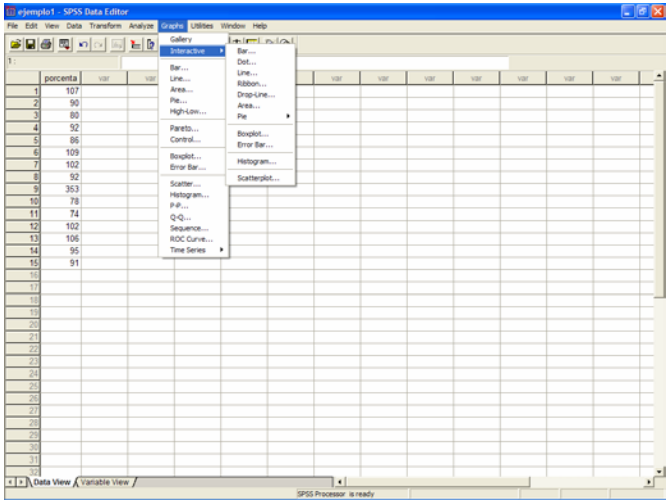
a) Generar un conjunto de datos en SPSS, indicando un **nombre de variable** (nota: sólo se consideran 8 caracteres,) y **tipo de dato** (en este caso numérico, de tamaño 3 con cero decimales).
b) Guardar los datos en un archivo (seguir la secuencia **FILE -> SAVE AS ->** en la caja de diálogo indicar la **unidad de disco**, **carpeta** y el **nombre del archivo** (nota: la extensión del archivo es **.SAV** y sólo se puede leer en SPSS).

Al guardar se tienen posibilidades de hacerlo en varios tipos, como: diferentes versiones de SPSS, en Excel, Lotus o dBase. Se recomienda guardar directamente en formato SPSS.

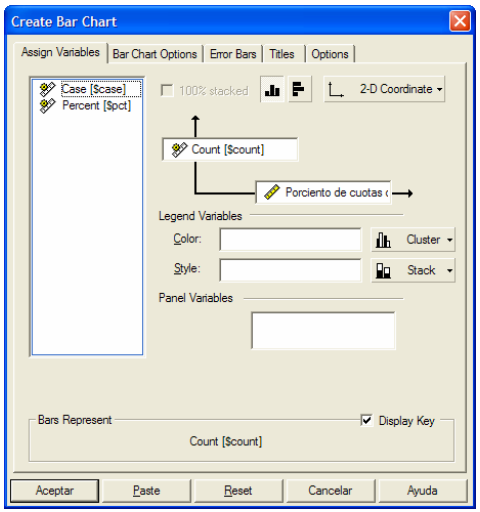


Análisis iniciales

Para empezar por las posibilidades gráficas, seleccionar del menú la opción **GRAPHS -> INTERACTIVE** y el **tipo de gráfico** a trabajar. Después sólo hay que leer y seleccionar las opciones que se presentan en las cajas de diálogo. Los usuarios avanzados y avezados pueden evitar la opción **interactive** e ir directamente al tipo de gráfico que requieren.



Crear un gráfico de barras



← ←

En los gráficos de barras, para definir qué variable corresponde a cada eje del gráfico basta con seleccionarlas de la lista de variables y hacer un simple "arrastre", además se tiene una amplia gama de posibilidades a la cual se tiene acceso mediante las opciones tipo fólder que se encuentran en la parte superior de la caja de diálogo.

Se deben resaltar las posibilidades de darle una orientación vertical u horizontal al gráfico, así como un aspecto en dos o tres dimensiones; con sólo dar un clic sobre el botón correspondiente.

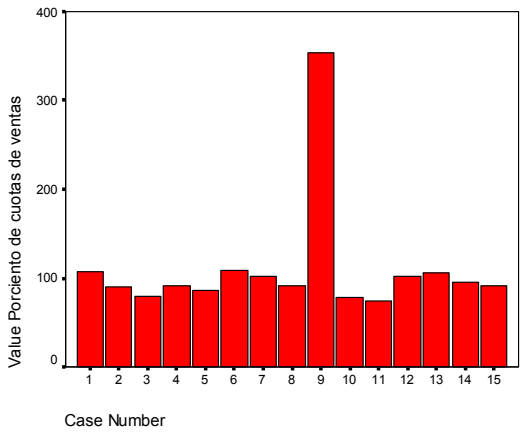
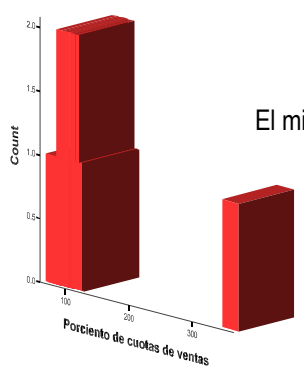


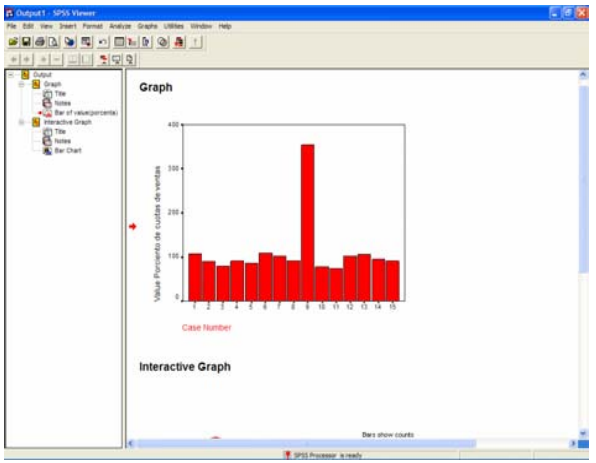
Gráfico de barras en dos dimensiones de los datos del ejemplo 1.



El mismo gráfico con aspecto de tres dimensiones.

La recomendación es explorar, mediante las opciones tipo fólder, las posibilidades de aspecto, colores y textos en el gráfico.

Al realizar el gráfico se despliega una ventana que muestra los resultados del análisis (**SPSS Viewer**).



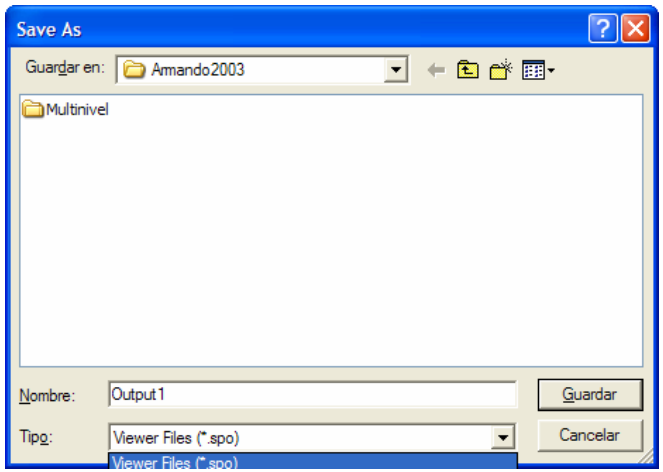
En el lado izquierdo aparece la lista de los procedimientos realizados, y con seleccionar alguno de ellos mediante un clic se pueden ver los resultados correspondientes.

Es importante notar que los títulos y gráficos se pueden modificar (editar), para esto basta con dar un doble clic sobre ellos para seleccionarlos.

Para regresar de esta ventana a la de datos hay que ir a la opción **Window** del menú y seleccionar **SPSS Data Editor**. El mismo procedimiento funciona para ir del editor de datos a la ventana de resultados. También se puede seleccionar la ventana a trabajar de la barra de tareas, en la parte inferior de la pantalla.

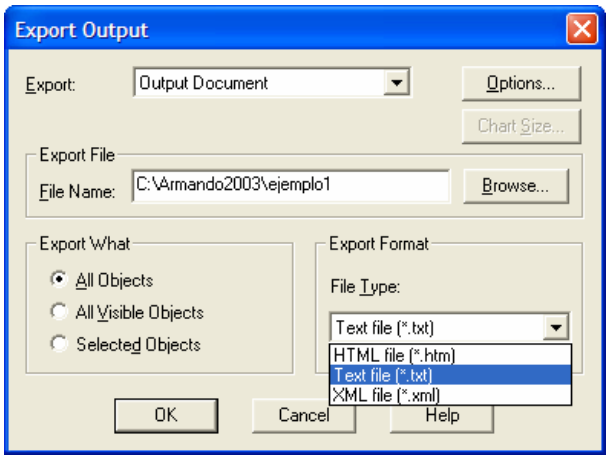
Guardar resultados

Para almacenar los resultados en disco hay que seguir la secuencia: **FILE -> SAVE AS** y en la caja de diálogo definir la **carpeta** donde guardar y el **nombre del archivo**.



Los resultados se almacenan como archivo con extensión **.SPO**, en un formato que sólo se puede consultar en SPSS.

Para leer estos resultados en cualquier procesador de palabras, la recomendación es exportarlos, con la secuencia de menú: **FILE -> EXPORT** y seleccionar opciones de la caja de diálogo.



Por ejemplo, en **Export** hay que definir si se exportan **todos los resultados**, nada más el **texto** o **sólo los gráficos**. Sin olvidar la carpeta, nombre y el tipo de archivo: HTML, XML o TEXTO.

El formato TEXT permite ver los resultados en cualquier editor o procesador de texto, como el bloc de notas o Word. Si por alguna razón el aspecto del archivo exportado no le agrada, la recomendación es:

Abrir de manera conjunta SPSS y el procesador de palabras de su preferencia (en nuestro caso Word) y realizar la secuencia.

En SPSS

- Seleccionar todo en la ventana de salida (**EDIT -> SELECT ALL**) o en su defecto seleccionar los procedimientos de su interés.
- Copiar todos los objetos (**COPY OBJECTS**)

En Word

- Ir al sitio del documento donde queremos insertar los resultados.
- Pegar y listo.

Se recomienda que toda la edición de gráficos, como colores, títulos y subtítulos se realice en SPSS. Para evitar problemas de memoria en la computadora o que el procesador de textos se “alente” o simple y sencillamente se “bloquee”.

Con esto se tienen los fundamentos para empezar a trabajar en SPSS, asegurándose de no “perder” los datos o resultados. Aunque la verdadera habilidad se obtiene a través de la práctica, así que a darse tiempo y armarse de paciencia para desentrañar los misterios y sorpresas que este paquete nos tenga reservados.

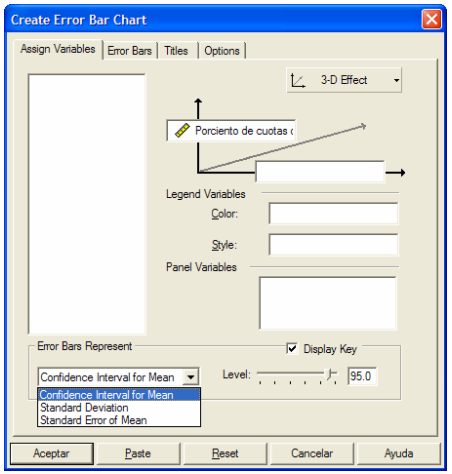
Capítulo 2

Describiendo los datos

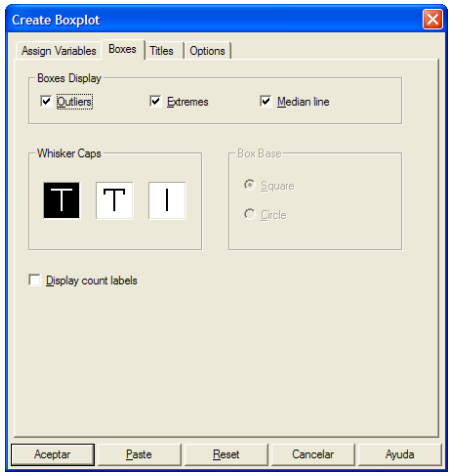
La primera actividad a realizar en un análisis estadístico es ver los datos, observarlos o explorarlos. Apreciar su distribución, agrupamiento, dispersión o la presencia de valores extremos. Como se dice: dejar que los datos hablen.

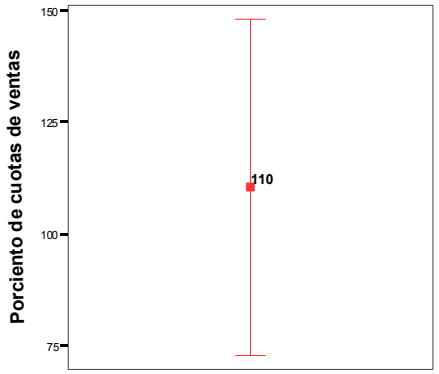
Dos de los gráficos más útiles, para este propósito, son el de barras con error y el de cajas con alambre (boxplot).

Gráficos de barras con error



Considerando los mismos datos del ejemplo 1, seleccionar del menú **GRAPHS -> INTERACTIVE -> ERROR BAR**, y en la caja de diálogo seleccionar la opción a desplegar en el gráfico: intervalos de confianza, desviación estándar o el error estándar de la media.



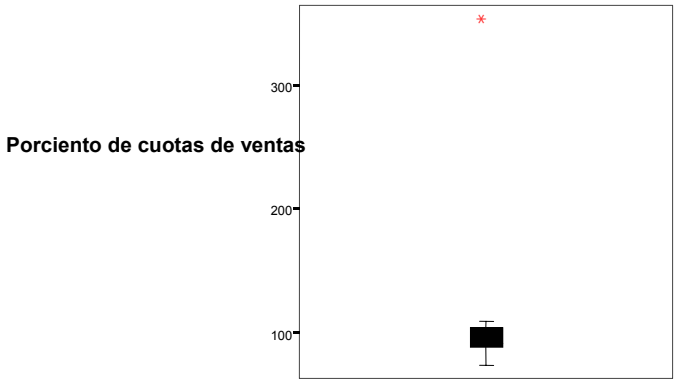


Resultado para la opción intervalo de confianza.

Gráfico de caja con alambre (bigote de gato o boxplot)

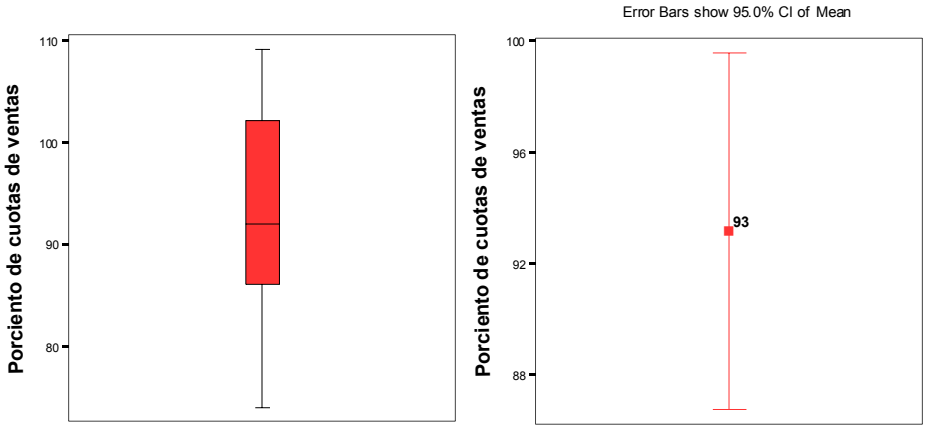
Con los mismos datos del ejemplo del capítulo anterior, seleccionar del menú **GRAPHS -> INTERACTIVE -> BOXPLOT**

De la caja de diálogo seleccionar las opciones correspondientes, sin olvidar las posibilidades del menú tipo fólder.



El gráfico de cajas y alambres muestra un valor aberrante, que definitivamente influye en el comportamiento de los datos. Desde el punto de vista práctico hay que revisar la veracidad del dato y considerar la pertinencia de eliminarlo del análisis o seguir trabajando con él. Desde el punto de vista numérico, mejor se elimina para no sesgar el análisis. Es importante recalcar que **ningún dato debe eliminarse sin una revisión y justificación previa**.

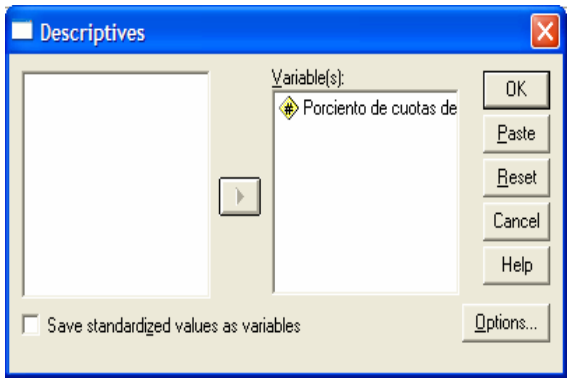
Al eliminar el valor 353 de los datos y rehacer los dos gráficos se tiene.



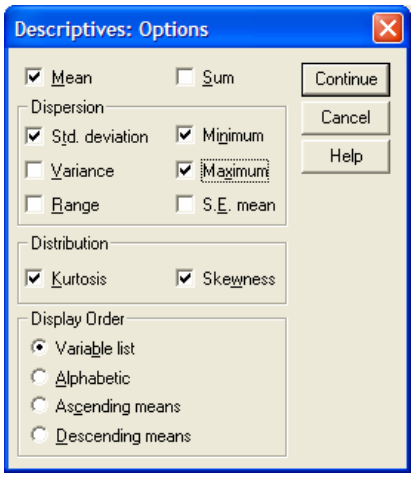
Donde se puede apreciar que el cambio de aspecto es más notorio en el gráfico de cajas. Una idea más clara de este cambio se aprecia al obtener las estadísticas descriptivas.

Para esto hay varias opciones, una de ellas es seguir la secuencia:

ANALYZE -> DESCRIPTIVE-STATISTICS -> DESCRIPTIVES



Selecionar la o las variables a trabajar.



Al presionar OPTIONS aparece una caja de diálogo donde se pueden seleccionar las estadísticas a calcular.

Las opciones Kurtosis y Skewness, proporcionan información sobre la distribución de los datos, por lo que se recomienda activarlas.

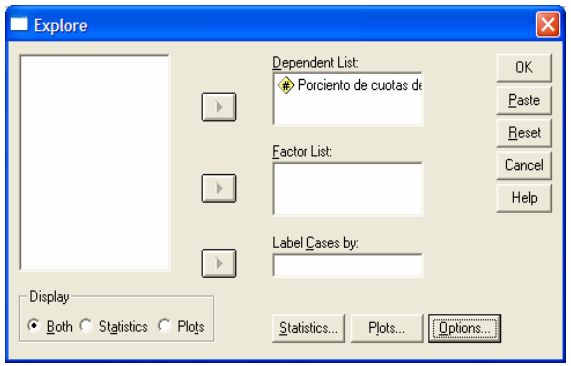
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Porciento de cuotas de ventas	15	74	353	110.47	67.944
Valid N (listwise)	15				

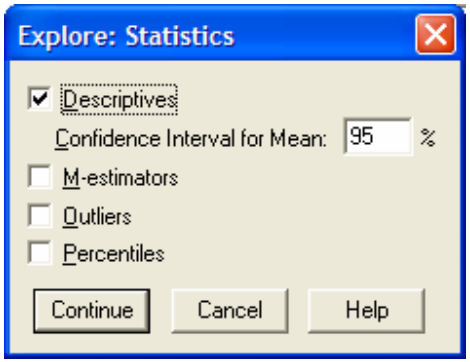
Otra opción, que aporta una mayor cantidad de información tanto gráfica como numérica es:

ANALYZE -> DESCRIPTIVE-STATISTICS -> EXPLORE

Cuya primera caja de diálogo es:



Aquí se pueden seleccionar las variables a trabajar, así como acceder a las opciones gráficas y estadísticas.



Por ejemplo, se puede definir cuales opciones se deben obtener o mostrar en los gráficos.

Los resultados son:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Porciento de cuotas de ventas	15	100.0%	0	.0%	15	100.0%

Descriptives

			Statistic	Std. Error
Porciento de cuotas de ventas	Mean		110.47	17.543
	95% Confidence Interval for Mean	Lower Bound	72.84	
		Upper Bound	148.09	
	5% Trimmed Mean		99.02	
	Median		92.00	
	Variance		4616.410	
	Std. Deviation		67.944	
	Minimum		74	
	Maximum		353	
	Range		279	
	Interquartile Range		20.00	
	Skewness		3.707	.580
	Kurtosis		14.097	1.121

Analizando algunos de estos resultados.

Skewness o Sesgo. Es una medida de la desviación de una muestra con respecto a la media central de una distribución normal. El sesgo es cero cuando se tiene una distribución simétrica con respecto a la media. Cuando es positivo indica que las observaciones se agrupan a la izquierda de la media, el signo del sesgo indica hacia que lado de la media se tienen los valores extremos.

Curtosis. Es una medida del pico o aplanado de una distribución normal. Una distribución normal estándar tiene una curtosis de 3. De tal manera que un valor mayor a 3 indica una distribución con pico mayor a una normal, mientras que un valor menor a 3 indica una distribución más aplanada que una normal

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average(Definition 1)	Porciento de cuotas de ventas	74.00	76.40	86.00	92.00	106.00	206.60	.
Tukey's Hinges	Porciento de cuotas de ventas			88.00	92.00	104.00		

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Porciento de cuotas de venta	.442	15	.000	.432	15	.000

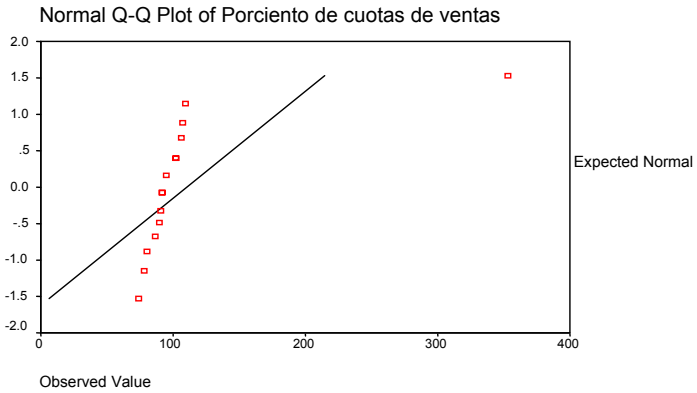
a. Lilliefors Significance Correction

La prueba de Kolmogorov tiene el siguiente juego de hipótesis.

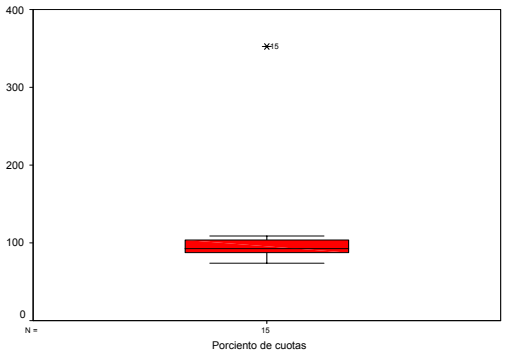
Ho: Los datos no tienen desviaciones significativas de una distribución normal (en otras palabras, se apegan a una distribución normal).
Ha: Los datos no se apegan a una distribución normal.

Para probar cualquier hipótesis es importante definir el juego de hipótesis a trabajar. Ya que la regla práctica de interpretación es: **rechazar Ho, si el valor de Sig. es menor a 0.05,** entonces se debe conocer a quien rechazar o no rechazar**.** En este caso Sig. = 0.000 por lo se tiene evidencia de que los datos no se apegan a una distribución normal.

```
Porciento de cuotas de ventas Stem-and-Leaf Plot

 Frequency     Stem &  Leaf

     2.00        7 .  48
     2.00        8 .  06
     5.00        9 .  01225
     5.00       10 .  22679
     1.00 Extremes    (>=353)

 Stem width:    10
 Each leaf:       1 case(s)
```



En este gráfico se confirma el resultado de la prueba de Kolmogorov-Smirnov, ya que los datos normales siguen la línea recta de referencia, lo que no sucede en este caso.

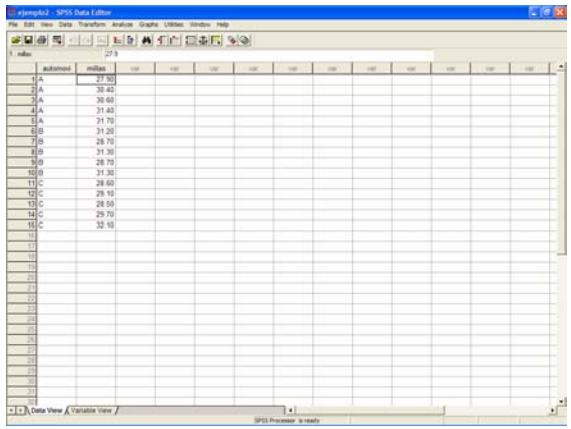


Los gráficos y estadísticas de una sola muestra dan información valiosa, pero hacen más sentido cuando se comparan dos o más muestras. Por lo que analizaremos el siguiente ejemplo.

Ejemplo. Un servicio de prueba de consumo obtuvo los siguientes resultados de milla por galón en cinco recorridos de prueba realizados con cada uno de tres automóviles compactos:

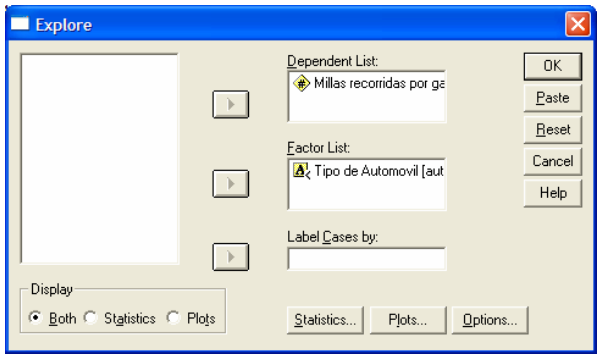
Automóvil A	27.9	30.4	30.6	31.4	31.7
Automóvil B	31.2	28.7	31.3	28.7	31.3
Automóvil C	28.6	29.1	28.5	32.1	29.7

(Estadística Elemental, John E. Freund y Gary A. Simon, 1992, Prentice Hall, pág. 56).



Primero hay que "darle" los datos a SPSS, nótese el formato en dos columnas: una que identifica el tipo de automóvil y otra para las millas recorridas.

Hay que arrastrar a la caja de diálogo correspondiente tanto la variable de respuesta o dependiente y el factor o variable independiente (variable que identifica el tipo de automóvil).



A continuación se presentan los resultados de **ANALYZE -> DESCRIPTIVE-STATISTICS -> EXPLORE**

Explore

Tipo de Automóvil

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	Tipo de Automovil	N	Percent	N	Percent	N	Percent
Millas recorridas por galón	A	5	100.0%	0	.0%	5	100.0%
	B	5	100.0%	0	.0%	5	100.0%
	C	5	100.0%	0	.0%	5	100.0%

Percentiles

			Percentiles						
		Tipo de Automovil	5	10	25	50	75	90	95
Weighted Average(Definition 1)	Millas recorridas por galón	A	27.9000	27.9000	29.1500	30.6000	31.5500	.	.
		B	28.7000	28.7000	28.7000	31.2000	31.3000	.	.
		C	28.5000	28.5000	28.5500	29.1000	30.9000	.	.
Tukey's Hinges	Millas recorridas por galón	A			30.4000	30.6000	31.4000		
		B			28.7000	31.2000	31.3000		
		C			28.6000	29.1000	29.7000		

Descriptives

	Tipo de Automovil			Statistic	Std. Error
Millas recorridas por galón	A	Mean		30.4000	.67007
		95% Confidence Interval for Mean	Lower Bound	28.5396	
			Upper Bound	32.2604	
		5% Trimmed Mean		30.4667	
		Median		30.6000	
		Variance		2.245	
		Std. Deviation		1.49833	
		Minimum		27.90	
		Maximum		31.70	
		Range		3.80	
		Interquartile Range		2.4000	
		Skewness		-1.538	.913
		Kurtosis		2.645	2.000
	B	Mean		30.2400	.62897
		95% Confidence Interval for Mean	Lower Bound	28.4937	
			Upper Bound	31.9863	
		5% Trimmed Mean		30.2667	
		Median		31.2000	
		Variance		1.978	
		Std. Deviation		1.40641	
		Minimum		28.70	
		Maximum		31.30	
		Range		2.60	
		Interquartile Range		2.6000	
		Skewness		-.605	.913
		Kurtosis		-3.328	2.000
	C	Mean		29.6000	.66030
		95% Confidence Interval for Mean	Lower Bound	27.7667	
			Upper Bound	31.4333	
		5% Trimmed Mean		29.5222	
		Median		29.1000	
		Variance		2.180	
		Std. Deviation		1.47648	
		Minimum		28.50	
		Maximum		32.10	
		Range		3.60	
		Interquartile Range		2.3500	
		Skewness		1.705	.913
		Kurtosis		2.939	2.000

Extreme Values[a]

	Tipo de Automovil			Case Number	Value
Millas recorridas por galón	A	Highest	1	5	31.70
			2	4	31.40
		Lowest	1	1	27.90
			2	2	30.40
	B	Highest	1	10	31.30
			2	8	31.30
		Lowest	1	7	28.70
			2	9	28.70
	C	Highest	1	15	32.10
			2	14	29.70
		Lowest	1	13	28.50
			2	11	28.60

a. The requested number of extreme values exceeds the number of data points. A smaller number of extremes is displayed.

En estos datos hace más sentido analizar si el rendimiento promedio es mayor en alguno de los tres tipos de automóvil, así como su variabilidad, obteniendo la mayor información posible de las estadísticas descriptivas.

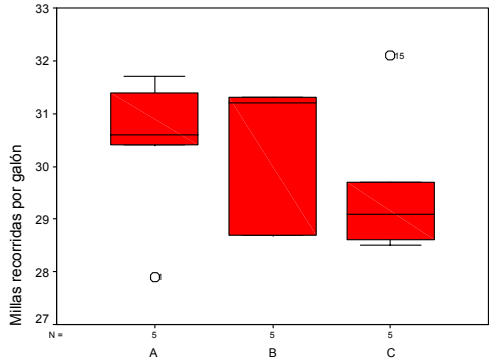
Millas recorridas por galón

Stem-and-Leaf Plots

```
Millas recorridas por galón

Stem-and-Leaf Plot for
AUTOMOVI= A

 Frequency    Stem &  Leaf
     1.00 Extremes    (=<27.9)
     1.00       30 .  4
     1.00       30 .  6
     1.00       31 .  4
     1.00       31 .  7

 Stem width:      1.00
 Each leaf:       1 case(s)
```

```
Millas recorridas por galón Stem-and-Leaf Plot for
AUTOMOVI= B

 Frequency    Stem &  Leaf
     2.00       28 .  77
      .00       29 .
      .00       30 .
     3.00       31 .  233

 Stem width:      1.00
 Each leaf:       1 case(s)
```

```
Millas recorridas por galón Stem-and-Leaf Plot for
AUTOMOVI= C

 Frequency    Stem &  Leaf
      .00       28 .
     2.00       28 .  56
     1.00       29 .  1
     1.00       29 .  7
     1.00 Extremes    (>=32.1)

 Stem width:      1.00
 Each leaf:       1 case(s)
```



Tipo de Automovil

Una buena cantidad del contenido de información se resume o concentra en este último gráfico, de donde se puede interpretar el rendimiento de los tres tipos de automóviles.

Al final de este capítulo se revisan algunos conceptos estadísticos y se da un glosario, como apoyo para entender e interpretar algunos de los resultados obtenidos hasta el momento.

Ejercicios. Nada más para ir "calentando dedo" y practicar lo que se ha visto hasta el momento.

Se tienen los siguientes datos de emisiones de óxido de azufre, en toneladas (modificado de Estadística Elemental, John E. Freund y Gary A. Simon, 1992, Prentice Hall, pp. 21-22).

Planta industrial A

15.8	22.7	26.8	19.1	18.5	14.4	8.3	25.9	26.4	9.8
22.7	15.2	23.0	29.6	21.9	10.5	17.3	6.2	18.0	22.9
24.6	19.4	12.3	15.9	11.2	14.7	20.5	26.6	20.1	17.0
22.3	27.5	23.9	17.5	11.0	20.4	16.2	20.8	13.3	18.1

Planta industrial B

27.8	29.1	23.9	24.4	21.0	27.3	14.8	20.9	21.7	15.8
18.5	22.2	10.7	25.5	22.3	12.4	16.9	31.6	22.4	24.6
16.5	27.6	23.0	27.1	12.0	20.6	19.7	19.9	26.5	21.4
28.7	23.1	16.2	26.7	13.7	22.0	17.5	21.1	34.8	31.5

a. "Teclear" y guardar los datos en un archivo SPSS (.SAV)
b. Realizar un análisis descriptivo de los datos
c. Comparar las dos plantas (sin realizar pruebas de hipótesis
d. Guardar los resultados del análisis en un archivo Word

Explore

Identificador de Planta

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	Identificador de Planta	N	Percent	N	Percent	N	Percent
Toneladas de Óxido de Azufre	A	40	100.0%	0	.0%	40	100.0%
	B	40	100.0%	0	.0%	40	100.0%

Descriptives

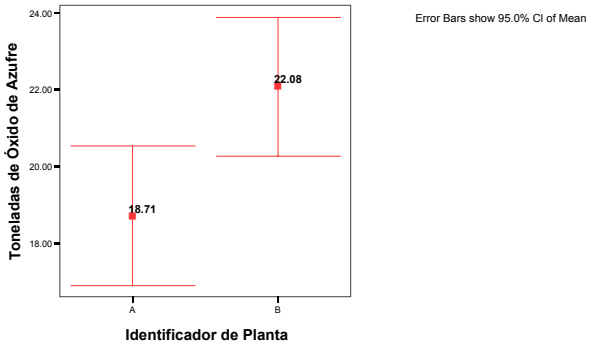
	Identificador de Planta			Statistic	Std. Error
Toneladas de Óxido de Azufre	A	Mean		18.7075	.90393
		95% Confidence Interval for Mean	Lower Bound	16.8791	
			Upper Bound	20.5359	
		5% Trimmed Mean		18.7972	
		Median		18.8000	
		Variance		32.684	
		Std. Deviation		5.71697	
		Minimum		6.20	
		Maximum		29.60	
		Range		23.40	
		Interquartile Range		8.0250	
		Skewness		-.196	.374
		Kurtosis		-.585	.733
	B	Mean		22.0850	.89519
		95% Confidence Interval for Mean	Lower Bound	20.2743	
			Upper Bound	23.8957	
		5% Trimmed Mean		22.0639	
		Median		22.1000	
		Variance		32.055	
		Std. Deviation		5.66168	
		Minimum		10.70	
		Maximum		34.80	
		Range		24.10	
		Interquartile Range		8.9000	
		Skewness		-.008	.374
		Kurtosis		-.341	.733

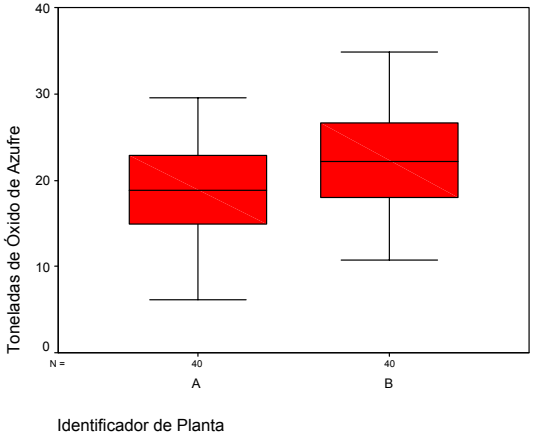
Toneladas de Óxido de Azufre

Stem-and-Leaf Plots

```
Toneladas de Óxido de Azufre Stem-and-Leaf Plot for
PLANTA= A

 Frequency     Stem &  Leaf
     3.00        0 .  689
     7.00        1 .  0112344
    12.00        1 .  555677788899
    12.00        2 .  000012222334
     6.00        2 .  566679
 Stem width:     10.00
 Each leaf:       1 case(s)
Toneladas de Óxido de Azufre Stem-and-Leaf Plot for
PLANTA= B

 Frequency     Stem &  Leaf
     5.00        1 .  02234
     8.00        1 .  56667899
    15.00        2 .  001111222233344
     9.00        2 .  566777789
     3.00        3 .  114
```

```
Stem width:     10.00
Each leaf:       1 case(s)
```





En este último gráfico se puede apreciar que la emisión de azufre es, en promedio, semejante en ambas plantas (tanto en la media como en dispersión). Aunque la planta B tiende a presentar una mayor emisión de Azufre.

Para entender e interpretar los resultados obtenidos hasta el momento, se revisan algunos conceptos estadísticos muy sencillos pero útiles.

TIPO DE DATOS Y NIVELES DE MEDICIÓN

Los datos pueden ser cualitativos o cuantitativos. Los **datos cualitativos**, como color de ojos en grupo de individuos, no se pueden trabajar aritméticamente, ya que son etiquetas que definen si un individuo pertenece o no a una categoría. Inclusive los datos de este tipo también se conocen como categóricos.

Datos cuantitativos: Mediciones que toman valores numéricos, para los cuales descripciones como media y desviación estándar son significativos. Se pueden clasificar en discretos y continuos.

Datos discretos: Se colectan por **conteo**, por ejemplo el número de productos defectuosos en un lote de producción.

Datos continuos: Se colectan por **medición** y se expresan en una escala continua. Por ejemplo la estatura de las personas.

La primera actividad en el análisis estadístico es contar o medir, ya que los datos representan un modelo de la realidad basado en escalas numéricas y medibles. Los datos vienen en forma Nominal, Ordinal, Intervalo y de razón o cociente.

Los datos categóricos se pueden medir sobre una escala nominal u ordinal. Mientras que en los datos continuos se pueden medir en escala de intervalo o de razón. En la escala de intervalo el valor cero y las unidades de medición son arbitrarias, mientras que en la escala de razón la unidad de medición es arbitraria pero el cero es una característica natural.

Considerando que la mayoría de las pruebas estadísticas clásicas se basan en que los datos se apeguen a una distribución normal, se prefieren los datos en escalas de intervalo o de razón.

MEDIDAS DE TENDENCIA CENTRAL

Una forma de representar a un conjunto de datos es a través de una medida de localización de la tendencia central, entre las cuales se tiene la media, mediana y moda.

Media. La media aritmética o simplemente promedio se obtiene sumando todos los valores de una muestra y dividiendo entre el número de datos.

$$Media = \frac{\sum_{i=1}^{n} y_i}{n}$$

NOTA: Se considera el valor de Y, por corresponder a la variable de respuesta, que es la que se mide y para no confundir con X´s o variables independientes que en muchos estudios son factores o variables categóricas.

El cálculo de la media utiliza todas las observaciones, de manera que se vuelve sensible a valores extremos, ya que valores extremadamente grandes o pequeños "jalan" el resultado de cálculo hacia ellos. Aún así, es la medida de tendencia central más utilizada, por presentar propiedades matemáticas convenientes para el análisis estadístico inferencial.

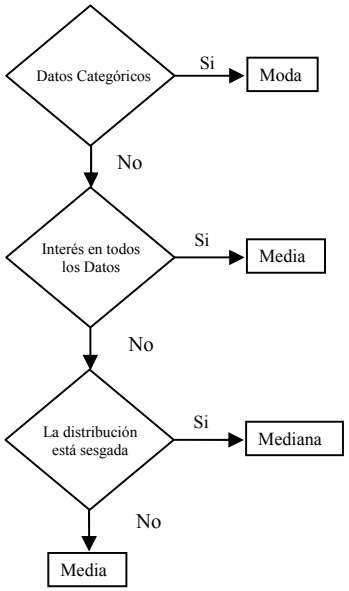
Media ponderada. Se utiliza cuando los datos de una muestra no tienen todos los mismos pesos, entonces cada valor se pondera por un peso acorde a su nivel de importancia.

Mediana. Es el valor que está exactamente a la mitad de un conjunto ordenado de observaciones. Si el número de observaciones es impar la mediana es un sólo valor. Si el número de observaciones es par entonces la mediana es el promedio de los dos valores que están al centro.

Generalmente la mediana es una mejor medida de tendencia central cuando hay valores extremadamente pequeños o grandes y cuando los datos están sesgados a la izquierda o la derecha.

Moda. La moda es el valor que más se presenta en un conjunto de observaciones.

Seleccionando entre media, mediana y moda



MEDIDAS DE DISPERSIÓN

Al hablar de dispersión se debe considerar que la calidad de la información y la variación están inversamente relacionadas. De aquí la necesidad de medir la variación que existe en un conjunto de datos.

Las medidas más comunes de variación son: el rango, varianza, desviación estándar y coeficiente de variación.

Rango. Es el valor absoluto de la diferencia del valor máximo menos el valor mínimo. Sólo se basa en dos valores y no es una medida recomendable cuando hay valores extremos.

Cuando se trabaja con números discretos es común definirlo como:

Valor máximo – Valor mínimo + 1.

Cuartiles. Cuando se tienen un conjunto de datos ordenados en forma ascendente, se pueden dividir en cuartos, Q_1, Q_2, Q_3 y Q_4. Para el valor del primer cuartil, Q_1, hay un 25% de valores más pequeños y un 75% de valores más grandes, de manera análoga en el Q_2=mediana hay un 50% de valores más pequeños y un 50% de valores más grandes y en Q_3 un 75% y 25% respectivamente.

Percentiles. Es un concepto semejante al de cuartil, de tal manera que $Q_1=P_{25}$, $Q_2=P_{50}$=mediana y $Q_3=P_{75}$. La ventaja de los percentiles es que pueden dividir a un conjunto de datos en 100 partes.

Rango intercuartilico. Expresa el intervalo de valores en el cual se encuentra el 50 % de los datos, ya que es la distancia del cuartil 1 al cuartil 3, esto es:

$$RIQ = Q_3 - Q_1$$

Varianza. Es un promedio de las distancias de cada observación con respecto a la media aritmética.

$$Varianza = S^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}$$

La varianza es una medida de la amplitud o dispersión de los datos y no está en las mismas unidades que las observaciones, de ahí que sea difícil su interpretación. Este problema se resuelve trabajando con la raíz cuadrada de la varianza.

Desviación estándar = S = $\sqrt{S^2}$

Coeficiente de variación. Expresa la variación de un conjunto de datos en relación a su media.

$$CV = 100\left|\frac{S}{\overline{Y}}\right|\%$$

El CV es independiente de las unidades de medición y en la estimación de un parámetro, cuando es menor al 10% el estimador se considera aceptable. Al inverso del CV, 1/CV, se le conoce como el cociente señal/ruido.

Para datos sesgados o agrupados, el coeficiente de variación cuartil es más útil que el CV.

$$V_Q = [0(Q_3-Q_1)/(Q_3+Q_1)]\%$$

SESGO Y CURTOSIS

Sesgo. Es una medida de la desviación de una muestra con respecto a la media central de una distribución normal. En otras palabras, mide la asimetría en la distribución de un conjunto de datos.

El sesgo es cero cuando se tiene una distribución simétrica con respecto a la media. Cuando es positivo indica que las observaciones se agrupan a la izquierda de la media, con la mayoría de los valores extremos a la derecha de la media. En otras palabras el signo del sesgo indica hacia que lado de la media se tienen los valores extremos.

$$\text{Sesgo (Skewness)} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^3}{(n-1)S^3}, \text{ con n > 1}$$

Curtosis. Es una medida del pico o aplanado de una distribución. Una distribución normal estándar tiene una curtosis de 3. De tal manera que un valor mayor que 3 indica un pico mayor a una distribución normal, mientras un valor menor que 3 indica una distribución más aplanada que una normal.

$$\text{Curtosis} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^4}{(n-1)S^4}, \text{ con n > 1}$$

OUTLIER

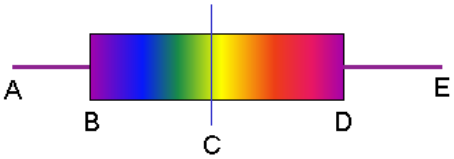
Un outlier u observación aberrante es un resultado distante de la mayoría de las observaciones. Se identifica porque su distancia al cuartil más cercano es mayor a 1.5 veces el rango intercuartílico.

BOX PLOT o diagrama de cajas y alambres

Estas gráficas se han vuelto muy populares, ya que ofrecen mucha información de manera compacta. Muestran el rango de los datos, la dispersión a través del rango intercuartílico y la mediana como medida de tendencia central. Recalcando que muestra la distribución de los datos.

Pasos para construir un BoxPlot

1. Calcular los cuartiles Q_1, Q_2 y Q_3
2. Sobre una línea, horizontal o vertical, "pintar": el valor mínimo; Q_1; Q_2; Q_3 y el valor máximo
3. Hacer un rectángulo de Q_1 a Q_3
4. Trazar una línea en Q_2=mediana
5. Revisar que los valores extremos no estén a una distancia mayor a 1.5 el valor del rango intercuartílico, si hay algún valor marcarlo con un asterisco.



En este gráfico, A=Q_1, B=Q_2 y C = Q_3.

GRÁFICO DE TALLO y HOJAS

Estos diagramas fueron desarrollados por John Tukey en 1977. Permiten observar la distribución de los datos originales y son muy útiles para resumir y describir, sobre todo cuando no se rebasan los cien datos.

Construir un diagrama de tallo y hoja:

1. Ordenar los datos (aunque este paso no es necesario si facilita su manejo)
2. Cada dato se divide en dos partes: el o los dígitos principales se convierten en el tallo y los dígitos posteriores en las hojas.

3. Colocar a la izquierda los dígitos más significativos del dato (Tallo), estos se escriben a lo largo del eje principal.
4. Colocar a la derecha los dígitos menos significativos (unidades o decimales), en orden de menor a mayor, para mostrar la distribución de los datos por cada dato se escribe una hoja. En algunos casos conviene poner en las hojas dos dígitos significativos.
5. Hacer un conteo de la frecuencia de valores asociados al valor del tallo.

La estadística, como toda disciplina científica, tiene su propio lenguaje o "jerga técnica" que permite y facilita la comunicación entre estadísticos o entre quien tiene los datos y quien debe analizarlos, de ahí la necesidad de definir y conocer algunos términos o conceptos.

GLOSARIO

Población: Colección de personas, animales, plantas o cosas acerca de las cuales se colectan datos. Es el grupo de interés sobre el cual se quieren realizar conclusiones.

Variables cualitativas y cuantitativas: Cualquier objeto o evento, que puede variar en observaciones sucesivas, ya sea en cantidad o cualidad. De aquí que se clasifiquen como cuantitativas o cualitativas, cuyos valores se denominan "variedades" y "atributos" respectivamente.

Variable: Una característica o fenómeno, que toma valores diferentes, como: peso o género, ya que difieren de medición a medición.

Aleatoriedad: esto significa impredecibilidad. Lo fascinante es que aunque cada observación aleatoria puede no ser predecible por si sola, colectivamente siguen un patrón predecible llamado su función de distribución. Lo que permite asociarle una probabilidad a la ocurrencia de cada resultado.

Muestra: Un subconjunto perfectamente acotado y definido de una población o universo.

Experimento: Es un proceso cuyos resultados no se conocen de antemano ni son predecibles.

Experimento estadístico: En general un experimento es una operación en la cual se seleccionan o fijan los valores de una variable (variable independiente) y se miden o cuantifican los valores de otra variable (variable dependiente). Entonces, un experimento estadístico es una operación en la cual se fijan los valores de la variable independiente y se toma una muestra aleatoria de una población para inferir los valores de la variable independiente.

Diseño de Experimentos: Es una herramienta para adquirir conocimiento acerca de un fenómeno o proceso. Este conocimiento se puede utilizar para ganar competitividad, acortar el ciclo de desarrollo de un producto o proponer nuevos productos o procesos que cumplan o excedan la expectación de un comprador.

Variable aleatoria: Es una función real (se le llama variable pero en realidad es una función) que asigna un valor numérico a cada evento simple. Estas variables son necesarias ya que no se pueden realizar operaciones algebraicas sobre resultados textuales, lo que permite obtener estadísticas, como promedios y varianzas. Además de que cualquier variable aleatoria tiene asociada una distribución de probabilidades.

Probabilidad: En términos simples es una medida de la posibilidad, se puede definir como el cociente del número de casos de interés que se presentan en un estudio entre el número de casos totales. Se utiliza para anticipar el tipo de distribución que sigue un conjunto de datos y asociarlos a un modelo probabilístico. Es importante hacer notar que los fenómenos aleatorios no son azarosos, ya que presentan un orden que sólo emerge cuando se describen un gran número de corridas (por ejemplo, al lanzar dos veces una moneda rara vez se obtiene un sol y una águila, pero si la lanza digamos unas diez mil veces, lo más seguro es que exista una clara tendencia a obtener la mitad de lanzamientos como sol y la otra mitad como águila). La descripción matemática de la variación es central a la estadística, ya que la probabilidad requerida para la inferencia está orientada hacia la distribución de los datos y no es de ninguna manera axiomática o combinatoria.

Unidad Muestral: Es una persona, animal, planta o cosa que está bajo observación o estudio por un investigador. En otras palabras, el objeto o

entidad básica sobre la cual se realiza un experimento, por ejemplo, una persona, una muestra de suelo, etcétera.

Parámetro: Un valor desconocido, que por lo tanto tiene que estimarse. Los parámetros se utilizan para representar alguna característica de una población. Por ejemplo, la media poblacional, μ, es un parámetro que generalmente se utiliza para indicar el promedio de una cantidad.

En una población, un parámetro es un valor fijo que no cambia. Cada muestra de la población tiene su propio valor de alguna estadística que sirve para estimar su parámetro.

Estadística o estadístico: Valor que se obtiene a partir de una muestra de datos. Se utiliza para generar información acerca del parámetro de su población. Ejemplo de estadística es la media y la varianza de una muestra, que se representan con letras latinas, mientras que los parámetros de una población se representan con letras griegas.

Estadística descriptiva: Área de la estadística que permite presentar los datos de manera clara y concisa, de tal forma que para la toma de decisiones se tengan a la mano las características esenciales de los datos.

Los principales estadísticos como medidas de tendencia central son la media o la mediana y la varianza o la desviación estándar como medidas de dispersión.

Estadística inferencial. Implica o significa hacer inferencias de una población, partiendo de valores muestrales. Cualquier conclusión de este tipo se debe expresar en términos probabilísticos, ya que la probabilidad es el lenguaje y la herramienta de medición de la incertidumbre que cuantifica la validez de una conclusión estadística.

Inferencia estadística: Se refiere a incrementar el conocimiento de una población en base a datos muestrales, también se conoce como razonamiento inductivo e implica conocer el todo a partir de sus partes. La inferencia estadística guía la selección de un modelo estadístico apropiado que en combinación con el adecuado tipo de datos permite cuantificar la confiabilidad de las conclusiones que se obtienen.

Condiciones de Distribución normal: La distribución normal o Gaussiana es una distribución continua y simétrica que tiene la forma de una campana. Uno de los aspectos más interesantes es que sólo se necesitan la media y la varianza para determinar completamente la distribución. En estudios reales se han encontrado una amplia gama de variables que tienen una distribución aproximadamente normal. Y que aún cuando una variable puede seguir una distribución no-normal la media de muchas observaciones independientes de la misma distribución se acerca arbitrariamente a una distribución normal, conforme el número de observaciones aumenta.

Su mayor importancia radica en que muchas de las pruebas estadísticas más frecuentemente utilizadas tienen como condición que los datos sigan una distribución normal.

Estimación y prueba de hipótesis: La inferencia en estadística considera dos grandes temas: el primero es la estimación, que implica estimar valores para parámetros poblacionales, considerando la variación aleatoria, por lo que dichas estimaciones se dan en valores de intervalo y nunca como valores puntuales. El segundo gran tema corresponde al contraste o prueba de hipótesis, donde se someten a prueba los posibles valores de algún parámetro con base en un modelo estadístico probabilístico.

COMENTARIOS FINALES

Hasta el momento se han revisado los aspectos básicos del “paquete” SPSS y del análisis estadístico, por lo que estamos listos para empezar a revisar los fundamentos de la estadística inferencial, es decir estimaciones y contrastes o pruebas de hipótesis.

Capítulo 3

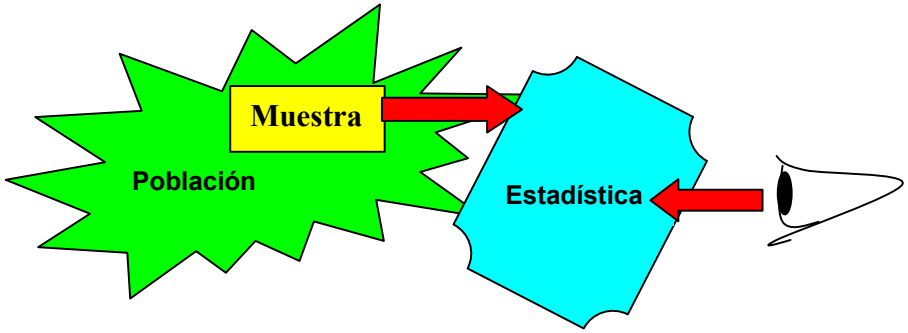
Introducción a la Inferencia

MOTIVACION

La estadística se caracteriza porque a través de una muestra se pueden realizar inferencias de toda una población en estudio. De manera que utilizando modelos estadísticos se puede asignar un nivel de confiabilidad a las conclusiones que se obtengan, proporcionando soporte para la toma de decisiones.

Población y muestra

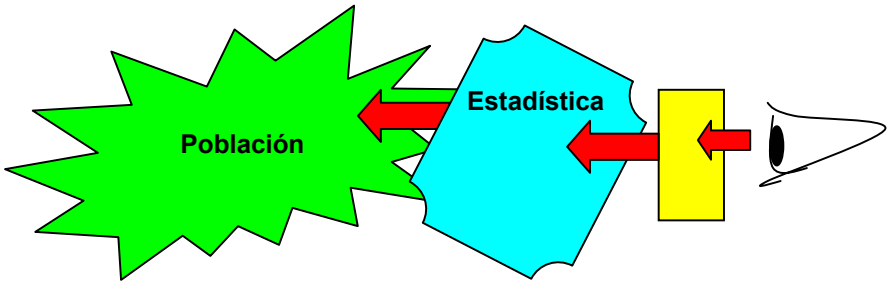
En cualquier proceso de investigación o producción es demasiado costoso, en recursos o en tiempo, revisar uno a uno todos los elementos que conforman una población, de ahí la necesidad de revisar unos cuantos, que sean representativos, y a partir de ellos predecir el comportamiento de toda la población.



El primer "viaje" a la estadística implica seleccionar una muestra de manera aleatoria, es decir, sin privilegiar o descartar de antemano elemento alguno; garantizando que todos tengan la misma posibilidad de ser elegidos. La mejor forma de hacer esto es utilizando herramientas como tablas de números aleatorios, una urna, o algún proceso de números pseudoaleatorios como los que vienen integrados en la mayoría de los paquetes estadísticos. Cualquiera de estas opciones es mejor que cerrar los ojos y estirar la mano o establecer criterios personales de selección de muestras.

Uno de los ejemplos más simples, pero nada estadístico, es lo que hacen quienes cocinan ya que a través de pequeñas "probadas" saben si un guiso está o no en su punto, esto previa homogenización del contenido de la cazuela y sin consumir todo su contenido.

Es conveniente aclarar que el tema de muestreo es una de las grandes ramas de la estadística, para la cual existen libros completos que analizan a detalles cada una de las opciones, dependiendo del propósito del muestreo.



El segundo "viaje" a la estadística consiste en analizar la muestra mediante alguna de las muchas técnicas de la estadística inferencial para tomar decisiones con respecto a la población, apoyándose en el conocimiento de causa evidenciado a partir de los datos y asignándole un nivel de confiabilidad o de incertidumbre a las conclusiones obtenidas.

INCERTIDUMBRE Y DISTRIBUCIONES ESTADÍSTICAS

La estadística es la disciplina que estudia los procesos estocásticos, es decir aquellos que presentan variaciones, sin causa asignable (debidas al azar). Por lo que se han desarrollado técnicas que permiten detectar y diferenciar variaciones por efecto de algún factor, de las debidas al azar, con el fin de identificar su comportamiento y reducir estas últimas a un nivel aceptable para que no altere las características de calidad de los productos en manufacturación.

Con el apoyo de la teoría de la probabilidad se ha demostrado que las variables aleatorias tienen un comportamiento bien definido, que se puede representar

mediante funciones de probabilidad y funciones de densidad de probabilidad, que dependiendo del tipo de unidades de medición generan las distribuciones estadísticas, base fundamental de las técnicas inferenciales. Debido a su importancia algunas de ellas se han tabulado para facilitar su uso; entre las más conocidas, sin ser las únicas, se encuentran:

- - Binomial

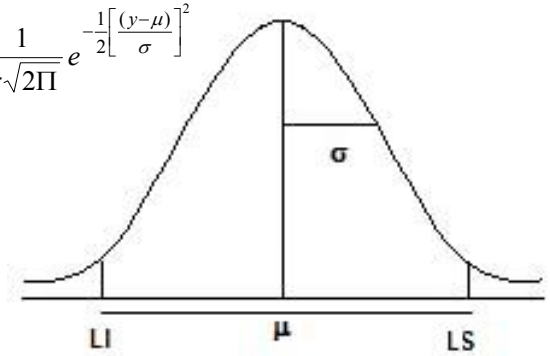
, variable que ya tabulada, esta distribución corresponde a valores de Z, con val

- Poisson

- Normal (Z)
  - t-student
  - F-Fisher
  - Ji-cuadrada (χ^2)

Estas distribuciones realmente corresponden a modelos matemáticos, por ejemplo la función de densidad de la distribución normal tiene como expresión matemática la siguiente ecuación.

$$f(y) = \frac{1}{\sigma\sqrt{2\Pi}} e^{-\frac{1}{2}\left[\frac{(y-\mu)}{\sigma}\right]^2}$$



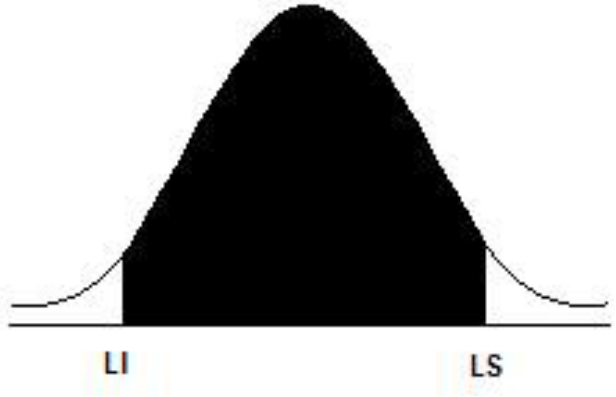
Distribución Normal, con intervalo de confianza para la media.

Donde se puede ver que la distribución queda totalmente representada por dos parámetros: μ (la media) y σ (la desviación estándar). Con las siguientes propiedades.

- Toda el área bajo la curva suma a 1.
- Los puntos de inflexión se localizan a más menos 1 desviación estándar.
- A más menos 4 desviaciones se encuentra la mayor parte del área bajo la curva (99.994%).

- A una distribución normal con $\mu = 0$ y $\sigma = 1$ se le conoce como normal estándar y se representa por la variable $z = \frac{(y-\mu)}{\sigma}$.

Cada conjunto de datos genera una distribución con sus propios valores de μ, σ y f(y), además es difícil que el valor estimado a partir de la media sea exactamente μ, por lo que es común establecer intervalos de confianza en los que se espera que el verdadero valor se encuentre entre un límite inferior (LI) y uno superior (LS). Valores que al representarse en la distribución, como área bajo la curva, indican una probabilidad.



Área bajo la curva delimitada por los límites de confianza.

Los valores de Z asociados a LI y LS acotan o delimitan cierta proporción del área, de ahí la importancia de saber, por ejemplo, que -1.96 < Z < 1.96 delimita el 95% del área bajo la curva de una distribución normal y que el área que no está sombreada corresponde al complemento a 1, en este caso al 5%, que expresado en probabilidades se le conoce como nivel de significancia, α, y a (1-α) como nivel de confianza.

De la misma forma el valor de -2.575 < Z < 2.575 delimita el 99%, con un complemento de 1% que dividido entre 2 corresponde al 0.5%($\alpha_{(0.01)/2}$=0.005), lo interesante es que al asociar estos valores a los datos muestrales se pueden establecer intervalos de confianza para estimar los valores poblacionales.

BREVE REVISIÓN A LA DISTRIBUCIÓN NORMAL Y TEOREMA DE LÍMITE CENTRAL

Distribución Normal

Definida por el modelo

$$f(y) = \frac{1}{\sigma\sqrt{2\Pi}} e^{-\frac{1}{2}\left[\frac{(y-\mu)}{\sigma}\right]^2}$$

Distribución normal estándar

$$f(y) = \frac{1}{\sigma\sqrt{2\Pi}} e^{-\frac{1}{2}[Z]^2}$$

Con **Z =** $\frac{(y-\mu)}{\sigma}$.

de la cual ya se habló al principio de este capítulo

Teorema del límite central

Este teorema establece que la distribución de las medias muestrales es normal aún cuando las muestras se toman de una distribución no-normal.

Si $x_1, x_2, \ldots, x_n$ son resultados de una muestra de n observaciones independientes de una variable aleatoria X con media μ_x y desviación σ_x, la media de las X´s se distribuirá aproximadamente como una distribución normal con media

$$\mu_{\bar{x}} \quad \text{y varianza} \quad \sigma^2_{\bar{X}} = \frac{\sigma^2_x}{n}$$

La aproximación es mucho mejor cuando n se hace grande. En general, la población de la cual se toman las muestras no necesita ser normal, para que la distribución de las medias muestrales sea normal. Esto constituye lo más notorio y poderoso de este teorema.

ESTIMACIÓN (INTERVALOS DE CONFIANZA)

La estimación hace referencia al cálculo de intervalos de confianza para los parámetros de una distribución, a partir de datos muestrales.

Por ejemplo, para la estimación de la media se tiene:

P{LI ≤ μ ≤ LS} = 1 - α

que puede leerse como: la probabilidad de que el verdadero valor de μ esté en el intervalo acotado por LI y LS es 1-α, cuyo resultado numérico es **LI ≤ μ ≤ LS**.

De aquí se pueden empezar a plantear las siguientes fórmulas de cálculo.

Parámetro	Intervalo
1) μ Con varianza conocida o **n > 30** (donde n es el tamaño de muestra).	$\bar{y} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ $y - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \le \mu \le y + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
2) μ Con varianza desconocida o **n ≤ 30**.	$\bar{y} \pm t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}$ $\bar{y} - t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}} \le \mu \le \bar{y} + t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}$
3) σ^2 Varianza distribución normal.	$\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}$

Intervalos de confianza para un parámetro.

El cuadro muestra los intervalos para los parámetros de una distribución normal: la media y la varianza. En la fórmula 1 se establece que la varianza es conocida, esto se logra cuando se tiene un proceso o fenómeno bien estudiado y se tiene una buena estimación del valor de la varianza poblacional. Cuando el tamaño de muestra es mayor a 30 se asume que $s^2 = \sigma^2$. En la fórmula 2 sólo se conoce la varianza muestral, así que para trabajar con ella hay que apoyarse en una distribución conocida como t de student, la cual también es simétrica y considera el manejo de n –1 grados de libertad. La fórmula 3 corresponde al intervalo para una varianza poblacional, a partir de la varianza muestral, aquí se debe utilizar una distribución conocida como Ji-cuadrada, y que se requieren dos valores uno para el límite inferior y otro para el límite superior, ya que esta

distribución no es simétrica y no tiene valores negativo, ya que al elevar al cuadrado un valor y luego sumarlo no hay posibilidades de obtener valores negativos.

4) $\mu_1 - \mu_2$ Con varianzas conocidas o **$n_1 > 30$** y **$n_2 > 30$.**

$$(\bar{y}_1 - \bar{y}_2) - Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le (\bar{y}_1 - \bar{y}_2) + Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

5) $\mu_1 - \mu_2$ Con varianzas desconocidas e iguales.

$$(\bar{y}_1 - \bar{y}_2) - t_{\frac{\alpha}{2},n_1+n_2-2} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \le (\bar{y}_1 - \bar{y}_2) + t_{\frac{\alpha}{2},n_1+n_2-2} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$ con

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}$$

Es importante recordar que se asume o supone $\sigma_1^2 = \sigma_2^2$ y que g.l. = n_1+n_2-2

6) $\mu_1 - \mu_2$ Con varianzas desconocidas y diferentes.

$$(\bar{y}_1 - \bar{y}_2) - t_{\frac{\alpha}{2},\nu}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \le \mu_1 - \mu_2 \le (\bar{y}_1 - \bar{y}_2) + t_{\frac{\alpha}{2},\nu}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

con

$$\nu = \text{grados de libertad} = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Se asume o supone $\sigma_1^2 \neq \sigma_2^2$

7) Razón o cociente de varianzas de dos poblaciones normales

$$\frac{S_1^2}{S_2^2} F_{1-\frac{\alpha}{2},n_1-1,n_2-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2} F_{\frac{\alpha}{2},n_1-1,n_2-1}$$

Es importante notar los dos valores de F, aunque si se obtiene uno el otro es su inverso, esto es: $F_{1-\alpha,u,v} = \frac{1}{F_{\alpha,v,u}}$

Intervalos de confianza para dos parámetros.

El cuadro 1 y 2 describen las fórmulas de cálculo para obtener intervalos de confianza para los parámetros de una población, considerando los valores muestrales (en este caso de la media y la varianza). Se debe aclarar que cada vez se utilizan menos, ya que el manejo numérico lo realiza el software, pero si es importante tenerlas en mente.

CONTRASTES O PRUEBAS DE HIPÓTESIS

Una hipótesis estadística es una aseveración acerca de los parámetros de una distribución de probabilidad.

Los procedimientos estadísticos de prueba de hipótesis se pueden utilizar para revisar la conformidad de los parámetros del proceso a sus valores especificados o para apoyar la modificación del proceso y lograr que se obtengan los valores deseados o especificados.

Para probar una hipótesis se toma una muestra aleatoria de la población en estudio, se calcula un estadístico de contraste adecuado, y se toma la decisión de rechazar o no rechazar la hipótesis nula Ho.

Ho. Hipótesis nula
Ha. Hipótesis alternativa

Al realizar un contraste de hipótesis se pueden cometer dos tipos de errores

α = P{error tipo I}
α = P{rechazar Ho/Ho es verdadera}
β = P{error tipo II}
β = P{no rechazar Ho/Ho es falsa}

Para α =0.05

Es importante notar que mientras más pequeño sea el valor de los extremos o colas de la distribución, se está más lejos de la zona de no rechazo de Ho.

Actualmente se hace referencia $\hat{\alpha}$ alfa gorra o el nivel de significancia estimado a partir de los datos, también identificado como P-value o Significancia. Este valor indica si la probabilidad del error tipo I es mayor o menor que el nivel preestablecido (0.05 o 0.01) y que con el uso del software

estadístico se ha vuelto fundamental para la interpretación de resultados. Reemplazando al uso de valores de tablas.

CONTRASTES DE HIPÓTESIS BASADOS EN LA DISTRIBUCIÓN NORMAL

VARIANZA(S) CONOCIDA(S)

8) Comparación de una media contra un valor definido por el investigador

Hipótesis	Estadístico de prueba	Regla de decisión Rechazar Ho si se cumple
Ho: $\mu = \mu_0$		$\lvert Z_c \rvert > Z_{\alpha/2}$
Ha: $\mu \neq \mu_0$		
	$Zc = \dfrac{\overline{Y} - \mu_0}{\dfrac{\sigma}{\sqrt{n}}}$	
Ho: $\mu \geq \mu_0$		$Z_c < -Z_\alpha$
Ha: $\mu < \mu_0$		
Ho: $\mu \leq \mu_0$		$Z_c > Z_\alpha$
Ha: $\mu > \mu_0$		

9) Comparación de un par de medias

Hipótesis	Estadístico de prueba	Regla de decisión Rechazar Ho si se cumple
Ho: $\mu_1 = \mu_2$		$\lvert Z_c \rvert > Z_{\alpha/2}$
Ha: $\mu 1 \neq \mu_2$		
	$Zc = \dfrac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	
Ho: $\mu_1 \geq \mu_2$		$Z_c < -Z_\alpha$
Ha: $\mu_1 < \mu_2$		
Ho: $\mu_1 \leq \mu_2$		$Z_c > Z_\alpha$
Ha: $\mu_1 > \mu_2$		

Al igual que en los intervalos de confianza, se supone varianza poblacional conocida, esto se logra cuando se tiene una buena estimación del valor de la varianza poblacional. Además, cuando el tamaño de muestra es mayor a 30 se asume que $S^2 = \sigma^2$.

VARIANZA(S) DESCONOCIDA(S)

10) Comparación de una media contra un valor definido por el investigador

Hipótesis	Estadístico de prueba	Regla de decisión Rechazar Ho si se cumple
Ho: $\mu = \mu_0$		$\lvert t_c \rvert > t_{\alpha/2,\, n-1}$
Ha: $\mu \neq \mu_0$		
	$t_c = \dfrac{\overline{Y} - \mu_0}{\dfrac{S}{\sqrt{n}}}$	
Ho: $\mu \geq \mu_0$		$t_c < -t_{\alpha,\, n-1}$
Ha: $\mu < \mu_0$		
Ho: $\mu \leq \mu_0$		$t_c > t_{\alpha,\, n-1}$
Ha: $\mu > \mu_0$		

11) Comparación de un par de medias

Hipótesis	Estadístico de prueba	Regla de decisión Rechazar Ho si se cumple
Ho: $\mu_1 = \mu_2$		$\lvert t_c \rvert > t_{\alpha/2,\, \nu}$
Ha: $\mu 1 \neq \mu_2$		
	$t_c = \dfrac{\overline{Y}_1 - \overline{Y}_2}{S_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$	
Ho: $\mu_1 \geq \mu_2$		$t_c < -t_{\alpha,\, \nu}$
Ha: $\mu_1 < \mu_2$		
Ho: $\mu_1 \leq \mu_2$		$t_c > t_{\alpha,\, \nu}$
Ha: $\mu_1 > \mu_2$		

Esta prueba corresponde a la comparación de dos medias, cuando las varianzas son iguales, en cuyo caso

$Sp = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$, con grados de libertad $\nu = n_1 + n_2 - 2$

12) Comparación de un par de medias, con varianzas desconocidas pero diferentes

Hipótesis	Estadístico de prueba	Regla de decisión Rechazar Ho si se cumple
Ho: $\mu_1 = \mu_2$ Ha: $\mu 1 \neq \mu_2$	$t_c = \dfrac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$	$\lvert t_c \rvert > t_{\alpha/2,\, \nu}$
Ho: $\mu_1 \geq \mu_2$ Ha: $\mu_1 < \mu_2$		$t_c < -t_{\alpha,\, \nu}$
Ho: $\mu_1 \leq \mu_2$ Ha: $\mu_1 > \mu_2$		$t_c > t_{\alpha,\, \nu}$

Los grados de libertad se obtienen con

$$\nu = \frac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Los contrastes 10 a12 corresponden a pruebas t-student, ideal para muestras pequeñas, n menor que 30, o cuando no se tiene un buen estimador de la varianza poblacional y sólo se cuenta con la varianza muestral. Se debe recalcar que **para comparar dos medias, se debe realizar previamente un análisis de comparación de dos varianzas**.

COMPARACIÓN DE DATOS PAREADOS

La análisis de datos pareados es útil para comparar datos de antes y después, sobre todo donde es difícil conseguir material experimental en condiciones iniciales homogéneas. Por ejemplo, en investigaciones médicas, ya sea con seres humanos o animales de laboratorio, donde la diferencia entre el estado final y las condiciones iniciales antes de un tratamiento es mejor medida que sólo la medición final.

Hipótesis	Estadístico de prueba	Regla de decisión Rechazar Ho si se cumple
Ho: $\mu_d = \Delta_0$ Ha: $\mu_d \neq \Delta_0$	$t_c = \dfrac{\overline{d} - \Delta_0}{\dfrac{S_d}{\sqrt{n}}}$	$t_c > t_{\alpha/2,\, n-1}$ $t_c < -t_{\alpha/2,\, n-1}$
Ho: $\mu_d \leq \Delta_0$ Ha: $\mu_d > \Delta_0$		$t_c > t_{\alpha,\, n-1}$
Ho: $\mu_d \geq \Delta_0$ Ha: $\sigma^2_1 < \sigma^2_2$		$t_c < -t_{\alpha,\, n-1}$

PRUEBAS PARA VARIANZAS

Comparación de una varianza contra un valor definido por el investigador

Hipótesis	Estadístico de prueba	Regla de decisión Rechazar Ho si se cumple
Ho: $\sigma^2 = \sigma^2_0$ Ha: $\sigma^2 \neq \sigma^2_0$	$\chi^2_c = \dfrac{(n-1)S^2}{\sigma_0^2}$	$\chi^2_c > \chi^2_{\alpha/2,\, n-1}$ $\chi^2_c < \chi^2_{1-\alpha/2,\, n-1}$
Ho: $\sigma^2 \geq \sigma^2_0$ Ha: $\sigma^2 < \sigma^2_0$		$\chi^2_c < \chi^2_{1-\alpha,\, n-1}$
Ho: $\sigma^2 \leq \sigma^2_0$ Ha: $\sigma^2 > \sigma^2_0$		$\chi^2_c > \chi^2_{\alpha,\, n-1}$

Comparación de un par de varianzas

Hipótesis	Estadístico de prueba	Regla de decisión Rechazar Ho si se cumple
Ho: $\sigma^2_1 = \sigma^2_2$ Ha: $\sigma^2_1 \neq \sigma^2_2$	$F_c = \frac{S_1^2}{S_2^2}$	$F_c > F_{\alpha/2,\, n1-1,\, n2-1}$ $F_c < F_{1-\alpha/2,\, n1-1,\, n2-1}$
Ho: $\sigma^2_1 \geq \sigma^2_2$ Ha: $\sigma^2_1 < \sigma^2_2$	$F_c = \frac{S_2^2}{S_1^2}$	$F_c > F_{\alpha,\, n2-1,\, n1-1}$
Ho: $\sigma^2_1 \leq \sigma^2_2$ Ha: $\sigma^2_1 > \sigma^2_2$	$F_c = \frac{S_1^2}{S_2^2}$	$F_c > F_{\alpha,\, n1-1,\, n2-1}$

Lo más interesante es: ¿Cómo determinar intervalos de confianza y contrastes de hipótesis utilizando SPSS? Aspecto que se muestra mediante los siguientes ejemplos.

Ejemplo 1. Se realizaron seis determinaciones del contenido de hidrógeno de un compuesto cuya composición teórica es del 9.55%, ¿Difiere el valor promedio del teórico?

%H: 9.17, 9.09, 9.14, 9.10, 9.13, 9.27

Las hipótesis correspondientes son: **Ho:** $\mu = 9.55$ **Ha:** $\mu \neq 9.55$

La secuencia de análisis es:

I. Ingresar los datos en una sola columna

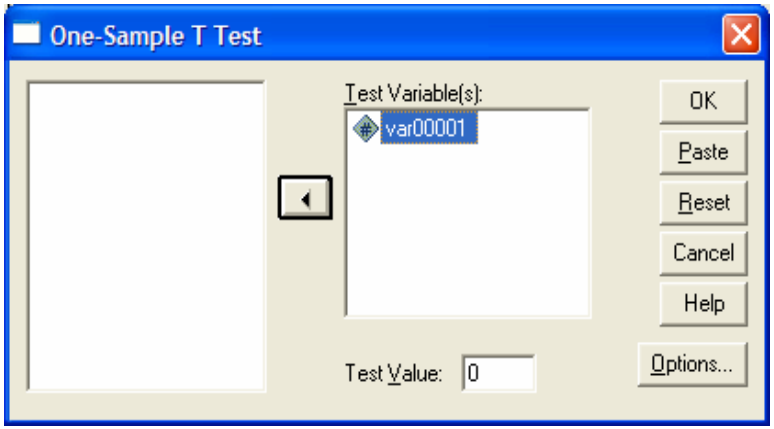
II. Realizar el análisis con la secuencia

ANALYZE -> COMPARE MEANS ->ONE-SAMPLE T TEST

III. Ingresar la variable de prueba (**Test Variable(s)**)

IV. Dar el valor de prueba, por cuestiones explicativas, en este ejemplo se dan los valores 0 y 9.55.

V. Para un mayor apoyo en el análisis se pide la opción explore, y sólo se salva el gráfico boxplot.



Resultados

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
VAR00001	6	9.1500	.06542	.02671

One-Sample Test

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference Lower	Upper
VAR00001	342.590	5	.000	9.1500	9.0813	9.2187

T-Test

One-Sample Statistics

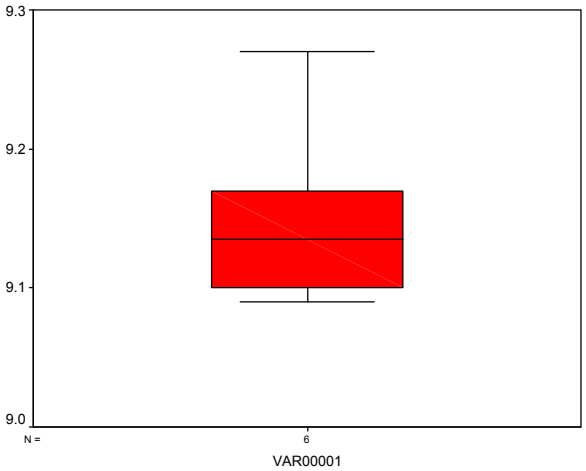
	N	Mean	Std. Deviation	Std. Error Mean
VAR00001	6	9.1500	.06542	.02671

One-Sample Test

	Test Value = 9.55					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
VAR00001	-14.977	5	.000	-.4000	-.4687	-.3313

ANÁLISIS:

1. Se tiene una media muestral de 9.15, con una desviación estándar muestral de 0.0654.
2. El intervalo de confianza al 95% va de 9.0813 hasta 9.2187, lo que permite ver que no contiene al valor de 9.55, y nos da evidencia de que el contenido de hidrógeno está por abajo de valor teórico.
3. Se comprueba esto con los valores de t = -14.98 que en colaboración con el de Sig. = 0.000, aportan evidencias estadística que permite rechazar Ho y afirmar con un 95% de confianza o con una significancia de 0.05 que el contenido de hidrógeno es diferente al 9.55%.



El gráfico boxplot (de cajas y alambres) tiene la siguiente información: una caja cuyos límites corresponden al cuartil 1 (Q_1) y el cuartil 3 (Q_3), la marca interior de la caja corresponde a la mediana y los extremos de los alambres corresponden a los valores mínimo y máximo. En este gráfico se puede ver que los datos están sesgados hacia valores pequeños, con un valor muy grande que no alcanza a equilibrar la caja del gráfico.

Ejemplo 2. Se analizó el contenido de silicio de una muestra de agua por dos métodos, uno de los cuales es una modificación del otro, en un intento por mejorar la precisión de la determinación. De acuerdo a los siguientes datos.

Método original(ppm)	Método modificado(ppm)
149	150
139	147
135	152
140	151
155	145

¿Es el método modificado más preciso que el regular?

Una medida de la precisión o dispersión está dada por la varianza, de manera que se pide una comparación de varianzas, con:

$$\textbf{Ho: } \sigma_1^2 \leq \sigma_2^2 \text{ vs } \textbf{Ha: } \sigma_1^2 > \sigma_2^2 .$$

El análisis se realiza con la secuencia:

I. Ingresar los datos en dos columnas: la **columna 1** identifica el método (Método), cuyas características son: numérica, label=tipo de método (valores 1=original y 2=modificado); la **columna 2** identifica la concentración de Silicio.

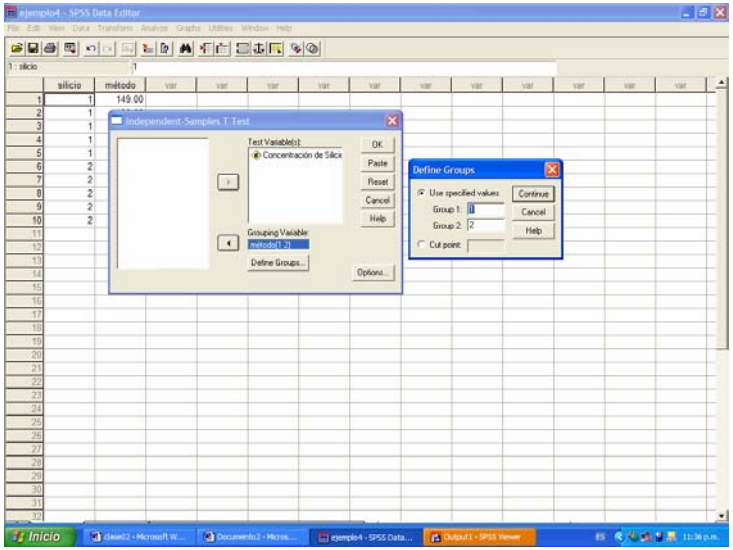
II. Seguir la secuencia:

ANALYZE -> COMPARE MEANS ->INDEPENDET-SAMPLES T TEST

III. Llenar la caja de diálogo que define **Test variable** (la variable de análisis, Silicio), la variable de agrupamiento (**grouping variable**, a variable que identifica los métodos. También aquí se debe definir los grupos, **Define groups**.

IV. dar OK y listo para que se realice el análisis

V. Para tener más herramientas de análisis se pide la opción explore, descrita en el capítulo anterior.



Resultados

Group Statistics

	Tipo de método	N	Mean	Std. Deviation	Std. Error Mean
Concentración de Silico en ppm	Método original	5	143.6000	8.17313	3.65513
	Método modificado	5	149.0000	2.91548	1.30384

Independent Samples Test

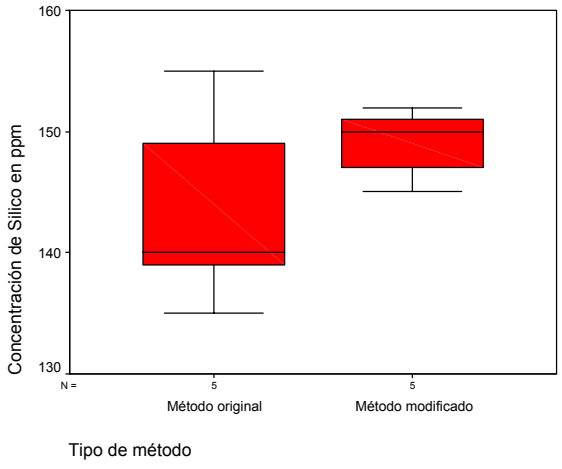
		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Concentración de Silico en ppm	Equal variances assumed	8.008	.022	-1.391	8	.202	-5.4000	3.88072	-14.34896	3.54896
	Equal variances not assumed			-1.391	5.002	.223	-5.4000	3.88072	-15.37467	4.57467

Descriptives

	Tipo de método			Statistic	Std. Error
Concentración de Silico en ppm	Método original	Mean		143.6000	3.65513
		95% Confidence Interval for Mean	Lower Bound	133.4517	
			Upper Bound	153.7483	
		5% Trimmed Mean		143.4444	
		Median		140.0000	
		Variance		66.800	
		Std. Deviation		8.17313	
		Minimum		135.00	
		Maximum		155.00	
		Range		20.00	
		Interquartile Range		15.0000	
		Skewness		.656	.913
		Kurtosis		-1.326	2.000
	Método modificado	Mean		149.0000	1.30384
		95% Confidence Interval for Mean	Lower Bound	145.3800	
			Upper Bound	152.6200	
		5% Trimmed Mean		149.0556	
		Median		150.0000	
		Variance		8.500	
		Std. Deviation		2.91548	
		Minimum		145.00	
		Maximum		152.00	
		Range		7.00	
		Interquartile Range		5.5000	
		Skewness		-.605	.913
		Kurtosis		-1.599	2.000

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Contenido de agua	Equal variances assumed	.088	.777	1.393	6	.213	.0260	.01867	-.01968	.07168
	Equal variances not assumed			1.290	3.448	.277	.0260	.02015	-.03366	.08566



ANÁLISIS:

La prueba de Levene muestra de que si hay diferencia en la precisión (o variabilidad) de un método a otro, lo que también se corrobora con los gráficos boxplot donde puede verse que hay menos variabilidad en el método modificado.

Es importante notar que en la prueba de muestras independientes aparecen dos opciones: Equal Variances Assumed o Equal Variances Not Assumed, **la decisión de cual utilizar depende del resultado de la prueba de Levene** (recordar que una prueba de t para comparar dos medias implica probar primero la igualdad de varianzas)

NOTA 1. Por el método clásico de F (cuyos resultados no se presentan aquí), el contraste de hipótesis de varianzas iguales no rechaza Ho.

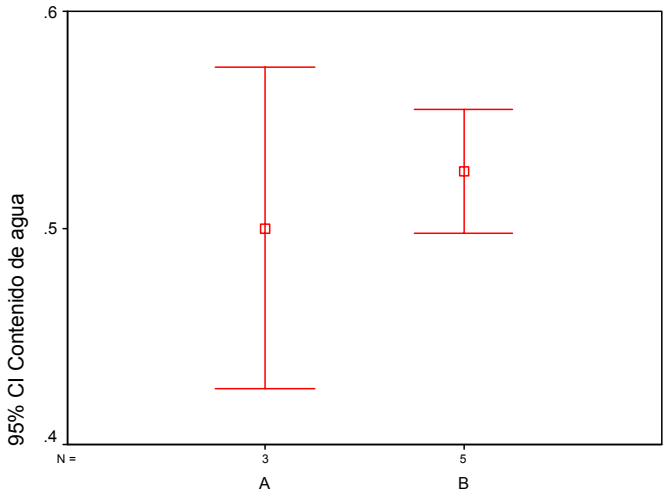
NOTA 2. La prueba de homogeneidad de varianzas de Levene es menos dependiente del supuesto de normalidad. Para cada caso se calcula el valor absoluto de la diferencia entre ese valor y la media, realizando un ANOVA de esas diferencias.

Ejemplo 3. Se analiza el contenido de agua en dos lotes de productos, por el método estándar de Karl Fischer. Con base en los datos del siguiente cuadro, **¿difieren los lotes en su contenido de agua?**

Contenido de agua	
Lote A	Lote B
0.50	0.53
0.53	0.56
0.47	0.51
	0.53
	0.50

La secuencia de procesamiento es igual a la del ejemplo anterior, pero ahora el énfasis está en los resultados de la media.

También se pide un gráfico de barras con error.



Group Statistics

	Tipo de lote	N	Mean	Std. Deviation	Std. Error Mean
Contenido de agua	B	5	.5260	.02302	.01030
	A	3	.5000	.03000	.01732

ANÁLISIS:

La hipótesis a trabajar (Ho:) es que la diferencia entre las medias es igual a cero, en otras palabras, las medias son iguales vs la Ha: de que la diferencia no es cero y por lo tanto las medias son diferentes.

La evidencia de la igualdad de medias se tiene en el gráfico del intervalo de confianza para cada media, donde para fines prácticos se puede considerar que si los intervalos se interceptan no hay evidencia estadística de diferencias entre las medias.

Lo recomendable es pedirle al "paquete" toda la información del ejemplo anterior y analizar con base en el resultado de la prueba de Levene y a la prueba de muestras independientes

Ejemplo 4. Se analiza un lote de productos para detectar concentraciones de hierro, antes y después de someterlos a un tratamiento para remover impurezas. De acuerdo a los siguientes datos, ¿hay evidencia de que el tratamiento es adecuado?

Este problema plantea una Ho: $\delta = 0$, donde δ es la diferencia de los valores entre antes y después del tratamiento, si la diferencia es cero entonces no hay efecto del tratamiento.

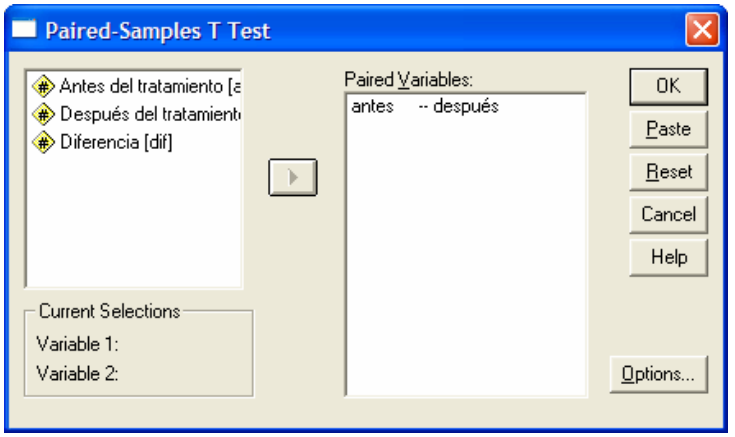
Lote A	Lote A Después del tratamiento
6.1	5.9
5.8	5.7
7.0	6.1
6.1	5.8
5.8	5.9
6.4	5.6
6.1	5.6
6.0	5.9
5.9	5.7
5.8	5.6

Secuencia de análisis

I. Primero y antes que nada se deben ingresar los datos en dos columnas, una para los valores de antes y otra para los valores de después.

II. Después se sigue la secuencia

ANALYZE -> COMPARE MEANS -> PAIRED-SAMPLES T TEST



En la caja de diálogo se definen las variables que forman el antes y el después.

Resultados

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Antes del tratamiento	6.1000	10	.36818	.11643
	Después del tratamiento	5.7800	10	.16865	.05333

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Antes del tratamiento & Después del tratamiento	10	.501	.140

Paired Samples Test

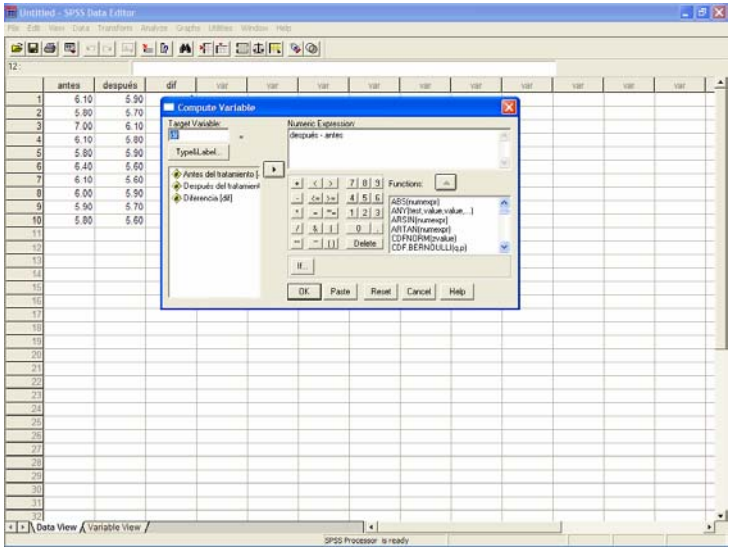
		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Antes del tratamiento - Después del tratamiento	.3200	.31903	.10088	.0918	.5482	3.172	9	.011

ANÁLISIS

Se puede ver en el intervalo de confianza para la diferencia que está por abajo de los valores de cero, corroborando con el valor de **Sig. =.011** que se puede rechazar la Ho.

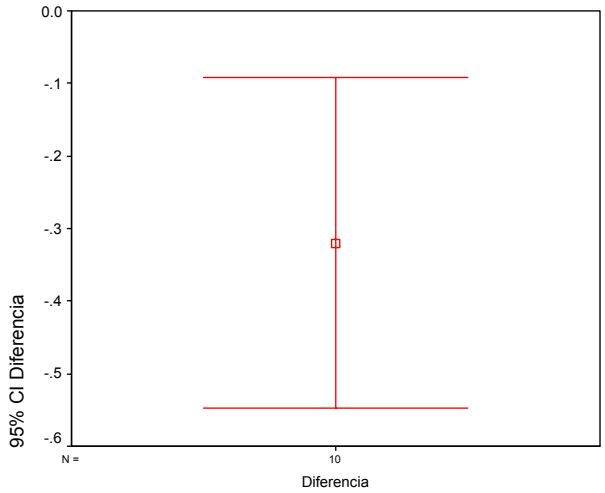
Para confirmar este análisis de manera visual y ensayar la generación de nuevas variables seguir la secuencia:

TRANSFORM -> COMPUTE



La caja de diálogo que aparece permite definir la nueva variables como se hace en una calculadora de bolsillo, en este caso dif =después- antes

. Con la nueva variable y practicando lo ya aprendido, obtener un gráfico de barras con error que muestre el intervalo de confianza para la diferencia.



Para seguir practicando y no aburrirse.

Ejercicio 1. Se analiza el contenido de agua en diez muestras de producción, comparando el método estándar de Karl Fischer y una versión coulombimétrica del método KF, **¿Hay evidencia de una diferencia real en los valores del contenido de agua?**

Coul KF	12.1	10.9	13.1	14.5	9.6	11.2	9.8	13.7	12.0	9.1
Regular KF	14.7	14.0	12.9	16.2	10.2	12.4	12.0	14.8	11.8	9.7

Ejercicio 2. Para motivar el ahorro de gasolina se planea una campaña a nivel nacional, pero antes de hacerlo a ese nivel se decide realizar un experimento para evaluar la eficiencia de dicha campaña. Para ello se realiza una campaña que promueve el ahorro de gasolina en un área geográfica pequeña, pero representativa. Se eligen al azar 12 familias y se mide la cantidad de gasolina utilizada un mes antes y un mes después de la campaña, obteniendo los siguientes datos.

(Estadística para las ciencias del comportamiento, 1998, Robert R. Pagano, International Thomson Editores, pág. 322)

Familia	Antes de la Campaña Galones/ mes	Después de la Campaña Galones/ mes
A	55	48
B	43	38
C	51	53
D	62	58
E	35	36
F	48	42
G	58	55
H	45	40
I	48	49
J	54	50
K	56	58
L	32	25

En estos dos ejercicios:

a) Guardar los datos en un archivo

b) Seleccionar la opción de análisis más adecuada, de acuerdo al enunciado del problema. PLANTEAR LAS HIPÓTESIS CORRESPONDIENTES.

c) Utilizar los apoyos gráficos que faciliten el análisis.

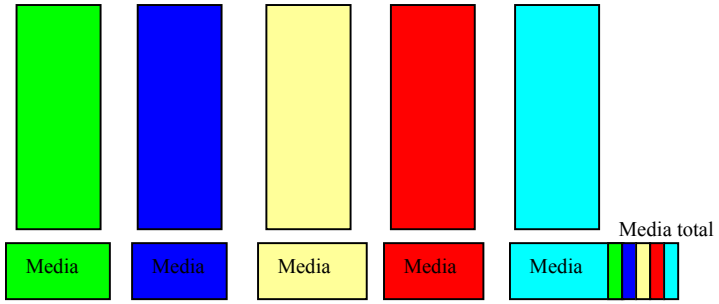
d) Interpretar los resultados y concluir al contexto del problema, todo esto en un archivo Word que incluya los resultados.

Capítulo 4

Análisis de Varianza y Diseño de Experimentos

MOTIVACIÓN AL ANÁLISIS DE VARIANZA (ANOVA, ANDEVA o ANVA)

Suponga un experimento donde se quieren comparar 5 tratamientos, para ver si su respuesta promedio es la misma para los 5 o si hay algunas diferentes.



De antemano el investigador asume que hay diferencia, si no que sentido tiene el experimento. También se sabe que en cada tratamiento debe haber un efecto de variaciones debida a la causa que se está controlando (temperatura, presión, etcétera) y una variación debida al azar, la cual es inevitable.

La variación entre tratamientos se mide como una varianza de la media de cada tratamiento con respecto a la gran media.

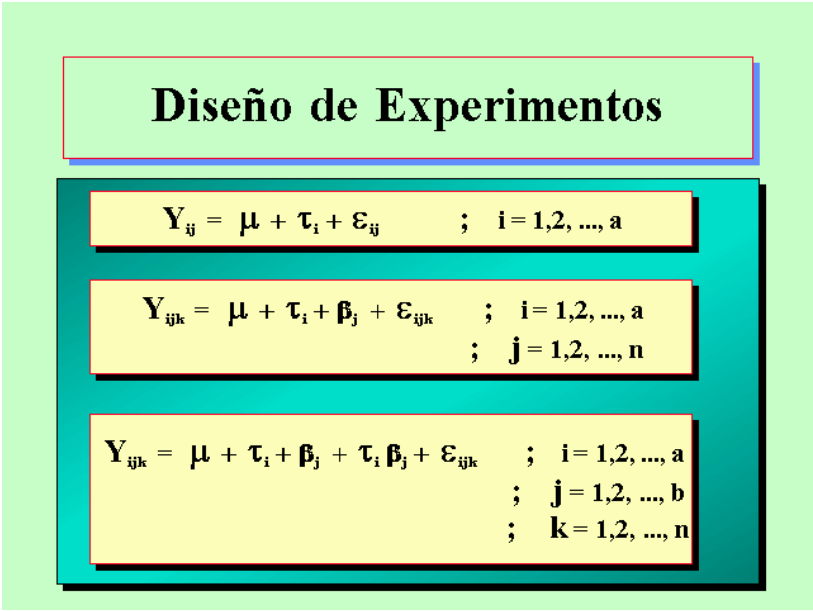
La variación dentro de tratamientos se mide comparando cada observación o medición con respecto a la media del respectivo tratamiento y en términos del análisis de varianza se le conoce como cuadrado medio del error.

Ahora, si se tienen dos varianzas lo que se puede hacer es compararlas mediante una prueba de F.

$$F = \frac{Varianza\_entre\_tratamientos}{Varianza\_dentro\_tratamientos}$$

La variación dentro de tratamientos se debe al azar y si no se puede establecer diferencia estadística entre estas varianzas, entonces no hay efecto de tratamiento y la variación se debe al azar.

MODELOS MÁS COMUNES EN EL DISEÑO DE EXPERIMENTOS



Diseño Completamente al Azar (DCA), de un factor o One-Way

La característica esencial es que todas las posibles fuentes de variación o de influencia están controladas y sólo hay efecto del factor en estudio. Este es el experimento ideal, todo controlado y lo único que influye es el factor de estudio.

Diseño de Bloques al Azar Completo (DBAC)

Sigue siendo un diseño de una vía pero hay alguna fuente con un gradiente de variación, que influye o afecta en el experimento, por lo tanto hay que cuantificar su efecto y eliminarlo de la varianza dentro de tratamientos, para evitar que nos conduzca a valores bajos de F y se llegue a conclusiones erróneas.

Diseños Factoriales

La tercera ecuación, de la figura anterior, muestra un diseño con dos factores de estudio, donde el mayor interés está en el efecto de la interacción, $\tau_i\beta_j$. Nótese la semejanza entre el modelo de la ecuación 2 y la 3.

ANÁLISIS DE VARIANZA DE UNA VÍA O DISEÑO COMPLETAMENTE AL AZAR.

Para mostrar los cálculos numéricos se tiene el siguiente ejemplo.

Un biólogo decide estudiar los efectos del etanol en el tiempo de sueño. Se seleccionó una muestra de 20 ratas, de edad semejante, a cada rata se le administró una inyección oral con una concentración en particular de etanol por peso corporal. El movimiento ocular rápido (REM) en el tiempo de sueño para cada rata se registró entonces durante un periodo de 24 horas, con los siguientes resultados.

(Modificado de Jay L. Devore, Probabilidad y estadística para ingeniería y ciencias, 5ª. Edición, Ed. Thomson Learning, México, 2001, pág. 412)

	0(control)	1 g/Kg	2 g/Kg	4 g/Kg	
	88.6	63	44.9	31	
	73.2	53.9	59.5	39.6	
	91.4	69.2	40.2	45.3	
	68	50.1	56.3	25.2	
	75.2	71.5	38.7	22.7	
$Y_{i.}$	**396.4**	**307.7**	**239.6**	**163.8**	$Y_{..}$ **= 1107.5**
$\bar{Y}_{i.}$	**79.28**	**61.54**	**47.92**	**32.76**	$\bar{Y}_{..}$ **= 55.375**

Como primer paso se debe tener en cuenta el par de hipótesis a trabajar.

Ho: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

Ha: $\mu_i \neq \mu_j$

Las fórmulas de cálculo son.

$$\sum_{i=1}^{a}\sum_{j=1}^{n}\left(Y_{ij}-\bar{Y}_{..}\right)^2 = n\sum_{i=1}^{a}\left(\bar{Y}_{i.}-\bar{Y}_{..}\right)^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}\left(Y_{ij}-\bar{Y}_{i.}\right)^2$$

A **B** **C**

Donde:

A es la suma de cuadrados Total
B Suma de cuadrados de tratamientos
C Suma de cuadrados del error

S.C.Total = $(88.6 - 55.375)^2 + (73.2 - 55.375)^2 + \ldots + (22.7 - 55.375)^2$
= 7 369.7575

S.C.Trat = $5[(79.28 - 55.375)^2 + (61.54 - 55.375)^2 + (47.92 - 55.375)^2 + (32.76 - 55.375)^2] = 5(1176.4715) = 5882.357$

S.C. Error = $(88.6 - 79.28)^2 + (73.2 - 79.28)^2 + \ldots + (22.7 - 32.76)^2 = 1487.4$

Tratamientos **a** = 4, con repeticiones **n** = 5, por lo tanto **a*n** = **N** = 4*5 = 20

Siguiendo otra estrategia de cálculo:

$\sum_{i=1}^{a}\sum_{j=1}^{n} Y_{ij}^2$ = $88.6^2 + 73.2^2 + 91.4^2 + 68^2 + \ldots + 45.3^2 + 25.2^2 + 22.7^2$ = 68 697.57

$\frac{Y_{..}^2}{N}$ = $(1107.5)^2/20$ = 61 327.8, valor que se conoce como factor de corrección

S.C.Total = $\sum_{i=1}^{a}\sum_{j=1}^{n} Y_{ij}^2 - \frac{Y_{..}^2}{N}$ = 68 697.57 – 61 327.8 = 7 369.77

S.C. Trat = $\frac{\sum_{i=1}^{a} Y_{i.}^2}{n_i} - \frac{Y_{..}^2}{N}$ = $(396.4^2 + 307.7^2 + 239.6^2 + 163.8^2)/5 - 61\,327.8$
= 336050.85/5 – 61327.8 = 5882.4

S.C.Error = S.C.Total – S.C.Trat = 7 369.77 – 5 882.4 = 1 487.4

Y la tabla de ANVA queda como:

Fuente de Variación	g.l.	Suma de Cuadrados	Cuadrados Medios	F	Pr > F
Tratamientos	3	5 882.4	1960.785	21.09	0.0001
Error	16	1 487.40	92.9625		
Total	19	7 369.757			

Donde se tiene evidencia para rechazar Ho, ya que Pr > F es mucho menor de 0.05.

DESPUÉS DEL ANÁLISIS DE VARIANZA

¿Cuál de todos los pares de medias son diferentes?

Para responder a esta pregunta se realizan pruebas de comparaciones múltiples de medias, como la de Tukey.

PRUEBA DE TUKEY

Este método se basa en utilizar el cuadrado medio del error, que se obtiene de un ANVA. Para calcular un valor ω que se compara con las diferencias de cada par de medias, si el resultado es mayor de ω se asumen medias diferentes en caso contrario se consideran semejantes o iguales.

La fórmula de cálculo es.

$$\omega = q_{\alpha}(a,\nu)\sqrt{\frac{CME}{n_g}}$$

donde:

a = número de tratamientos o niveles

ν = grados de libertad asociados al CME, con $\nu = n - a$

n_g = número de observaciones en cada uno de los a niveles

α = nivel de significancia

$q_{\alpha}(a,\nu)$ = valor crítico de rangos estudentizados (tablas)

La bibliografía reporta una amplia gama de pruebas, siendo las más comunes, además de la de Tukey, la de Fisher y la de Dunnet.

La prueba de Tukey y la de Fisher comparan todos los pares de medias, aunque Tukey genera intervalos más amplios que la de Fisher. Recomendando Tukey en estudios iniciales y la de Fisher en estudios finales o concluyentes.

La prueba de Dunnet permite comparar las medias contra un valor de referencia o control y dependiendo del "paquete" puede ser el primero o el último nivel del factor en estudio.

Después de comparar las medias, se recomienda verificar el cumplimiento de supuestos, para avalar la calidad de las conclusiones a las que se llega a través del análisis realizado: Homocedasticidad, Normalidad y comportamiento de residuales.

Homocedasticidad, varianzas homogéneas o significativamente iguales entre todos los tratamientos, aquí se recomienda la prueba de Barttlet

PRUEBA DE *BARTTLET* PARA HOMOGENEIDAD DE VARIANZAS

Esta prueba considera el siguiente par de hipótesis.

Ho: Todas las varianzas son iguales

Ha: Al menos dos varianzas son diferentes

Consiste básicamente en obtener un estadístico de contraste cuya distribución se aproxima a una distribución ji-cuadrada, con **a-1** grados de libertad, cuando las **a** muestras aleatorias son de poblaciones normales independientes. La secuencia de cálculo es.

1. Considerando la fórmula

$$\chi^2 = 2.3026\frac{q}{c}$$

2. Obtener $S_p^2 = \dfrac{\sum_{i=1}^{a}(n_i - 1)S_i^2}{N - a}$

3. Utilizar este resultado para calcular

$$q = (N-a)\log_{10} S_p^2 - \sum_{i=1}^{a}(n-1)\log_{10} S_i^2$$

4. Calcular $c = 1 + \dfrac{1}{3(a-1)}\left(\sum_{i=1}^{a}(n_i - 1)^{-1} - (N-a)^{-1}\right)$

5. Obtener el valor calculado de ji-cuadrada y compararlo con el valor de tablas con nivel de significancia α y **a-1** grados de libertad. Regla de decisión: Si $\chi^2_{calculada} > \chi_{\alpha,v}$, rechazar Ho.

PRUEBA DE *LEVENE* MODIFICADA

Debido a que la prueba de Bartlett es sensible al supuesto de normalidad, hay situaciones donde se recomienda un procedimiento alternativo, como lo es éste método robusto en cuanto a las desviaciones de la normalidad, ya que se basa en las medianas y no en las medias de los tratamientos. La secuencia de cálculo es:

Primero y antes que nada considerar el par de hipótesis a trabajar.

Ho: $\sigma_1^2 = \sigma_2^2 = ... = \sigma_a^2$ Todas las varianzas son iguales

Ha: Al menos dos varianzas son diferentes

1. Obtener la mediana de cada tratamiento: $\tilde{Y}_i$

2. Obtener para cada dato del experimento el valor absoluto de la desviación de cada observación con respecto a la mediana de su tratamiento. $d_{ij} = \left| Y_{ij} - \tilde{Y}_i \right|$

3. Sobre la tabla de estas diferencias, realizar un ANVA y aplicar la regla de decisión sobre el estadístico F para rechazar o no la Hipótesis nula.

PRUEBAS DE NORMALIDAD

Otra prueba consiste en verificar si los datos se comportan de acuerdo a una distribución normal, para lo cual existen pruebas numéricas y gráficas. Las numéricas básicamente plantean una curva normal teórica y mediante una prueba de falta de ajuste someten a prueba la hipótesis nula de que los datos se apegan a la distribución (Método de Kolmogorov-Smirnov, Anderson-Darling). Otro método es el gráfico, el cual es más utilizado por su impacto visual y lo fácil de su interpretación.

GRÁFICOS DE PROBABILIDAD NORMAL

Estos gráficos permiten juzgar hasta donde un conjunto de datos puede o no ser caracterizado por una distribución de probabilidad específica, en este caso la normal.

Gráficos de probabilidad acumulada.

Observación(i)	Y_i	Yi en orden ascendente	p_i (%)	Z_i	q_i
1	9.63	9.34	2.5	-1.96	-1.99
2	9.86	9.51	7.5	-1.44	-1.49
3	10.20	9.63	12.5	-1.15	-1.13
4	10.48	9.69	17.5	-0.94	-0.95
5	9.82	9.75	22.5	-0.76	-0.77
6	10.07	9.82	27.5	-0.60	-0.56
7	10.39	9.86	32.5	-0.46	-0.44
8	10.03	9.89	37.5	-0.32	-0.35
9	9.34	9.96	42.5	-0.19	-0.14
10	10.26	9.98	47.5	-0.06	-0.08
11	9.89	10.03	52.5	0.06	0.07
12	10.67	10.07	57.5	0.19	0.19
13	9.69	10.13	62.5	0.32	0.37
14	10.15	10.15	67.5	0.46	0.43
15	10.32	10.20	72.5	0.66	0.58
16	9.98	10.26	77.5	0.76	0.76
17	9.51	10.32	82.5	0.94	0.94
18	10.13	10.39	87.5	1.15	1.15
19	9.96	10.48	92.5	1.44	1.42
20	9.75	10.67	97.5	1.96	1.98

$$p_i = \frac{100(i-0.5)}{n}$$

$$q_i = \frac{Yi - \overline{Y}}{Sy}$$

Un gráfico de los pares (Y_i, p_i) se espera que tenga una forma de S para asegurar una aproximación normal, aunque es más común hacer este gráfico en papel normal para obtener una línea recta.

Si todos los puntos de los datos aparecen aleatoriamente distribuidos a lo largo de la línea recta y si la línea pasa sobre o cercanamente a la intersección de la media de Y, el 50% de probabilidad, el ajuste de los datos a la distribución normal se considera adecuado.

Contrariamente, si los puntos aparecen con forma de S, la sugerencia es que los datos no se distribuyen normalmente.

Con la ayuda de una tabla de probabilidad normal, las probabilidades acumuladas, p_i pueden convertirse en sus correspondientes valores normales estandarizados z_i.

$$P(Z \leq Z_i) = p_i$$

Si se conoce la media, la varianza y la variable Y_i, los datos muestreados se pueden estandarizar utilizando la transformación:

$$q_i = \frac{Y_i - \mu_y}{\sigma_y}$$

dado que la μ_y y σ_y generalmente no se conocen, se usa la ecuación:

$$q_i = \frac{Y_i - \bar{Y}}{S_y}$$

A continuación se puede hacer un gráfico de los puntos (q_i, z_i) que sirve para juzgar la normalidad de un conjunto de datos.

Si se traza una gráfica con la misma escala para q_i y Z_i, se espera que los puntos se distribuyan aleatoriamente a lo largo de una línea recta dibujada a 45º.

EJEMPLOS DEL ANÁLISIS DE VARIANZA

Ejemplo 1. Un fabricante supone que existe diferencia en el contenido de calcio en lotes de materia prima que le son suministrados por su proveedor. Actualmente hay una gran cantidad de lotes en la bodega. Cinco de estos son elegidos aleatoriamente. Un químico realiza cinco pruebas sobre cada lote y obtiene los siguientes resultados.

Lote 1	Lote 2	Lote 3	Lote 4	Lote 5
23.46	23.59	23.51	23.28	23.29
23.48	23.46	23.64	23.40	23.46
23.56	23.42	23.46	23.37	23.37
23.39	23.49	23.52	23.46	23.32
23.40	23.50	23.49	23.39	23.38

¿Hay variación significativa en el contenido de calcio de un lote a otro?

Las hipótesis a contrastar son:

Ho: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

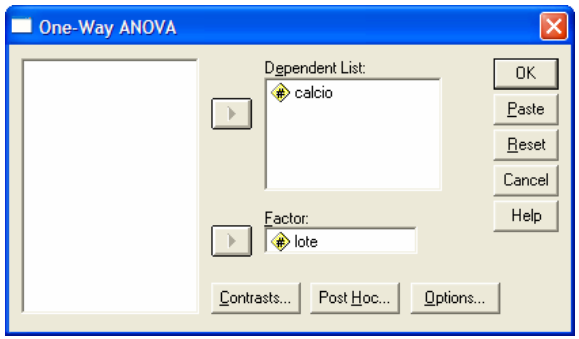
Ha: al menos un par de medias es diferente

El análisis se realiza con la secuencia:

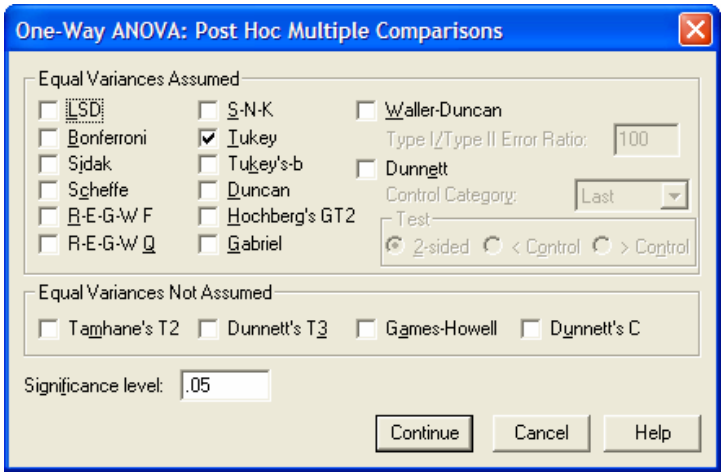
1. Ingresar los datos en dos columnas, una para identificar el lote y otra para la concentración de calcio.
2. Del menú seguir los pasos:

   ANALYZE -> COMPARE MEANS -> ONEWAY ANOVA

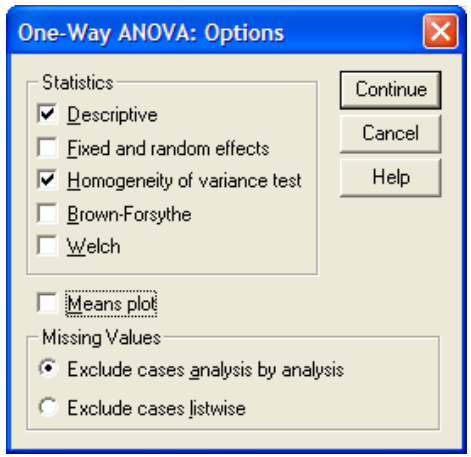
3. De la caja de diálogo seleccionar las variables a trabajar y colocarlas en su caja correspondiente.



4. Abrir el diálogo **Post Hoc** y seleccionar la prueba de comparaciones múltiples de medias.



5. Abrir el diálogo **options** y seleccionar las opciones a trabajar.



6. Ya con todas las opciones seleccionadas, dar **OK** para realizar el análisis.

RESULTADOS

Descriptives

CALCIO

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	5	23.4580	.06870	.03072	23.3727	23.5433	23.39	23.56
2	5	23.4920	.06301	.02818	23.4138	23.5702	23.42	23.59
3	5	23.5240	.06877	.03076	23.4386	23.6094	23.46	23.64
4	5	23.3800	.06519	.02915	23.2991	23.4609	23.28	23.46
5	5	23.3640	.06504	.02909	23.2832	23.4448	23.29	23.46
Total	25	23.4436	.08770	.01754	23.4074	23.4798	23.28	23.64

La primera tabla despliega estadísticas descriptivas, como la media, desviación estándar y los intervalos de confianza para los promedios de cada uno de los tratamientos.

ANOVA

CALCIO

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.097	4	.024	5.535	.004
Within Groups	.088	20	.004		
Total	.185	24			

ANVA con valores de F y Significancia

CALCIO

Tukey HSD[a]

LOTE	N	Subset for alpha = .05		
		1	2	3
5	5	23.3640		
4	5	23.3800	23.3800	
1	5	23.4580	23.4580	23.4580
2	5		23.4920	23.4920
3	5			23.5240
Sig.		.204	.094	.528

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 5.000.

Resultado de la comparaciones múltiples de medias.

Multiple Comparisons

Dependent Variable: CALCIO
Tukey HSD

(I) LOTE	(J) LOTE	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-.0340	.04186	.924	-.1593	.0913
	3	-.0660	.04186	.528	-.1913	.0593
	4	.0780	.04186	.368	-.0473	.2033
	5	.0940	.04186	.204	-.0313	.2193
2	1	.0340	.04186	.924	-.0913	.1593
	3	-.0320	.04186	.938	-.1573	.0933
	4	.1120	.04186	.094	-.0133	.2373
	5	.1280*	.04186	.044	.0027	.2533
3	1	.0660	.04186	.528	-.0593	.1913
	2	.0320	.04186	.938	-.0933	.1573
	4	.1440*	.04186	.019	.0187	.2693
	5	.1600*	.04186	.008	.0347	.2853
4	1	-.0780	.04186	.368	-.2033	.0473
	2	-.1120	.04186	.094	-.2373	.0133
	3	-.1440*	.04186	.019	-.2693	-.0187
	5	.0160	.04186	.995	-.1093	.1413
5	1	-.0940	.04186	.204	-.2193	.0313
	2	-.1280*	.04186	.044	-.2533	-.0027
	3	-.1600*	.04186	.008	-.2853	-.0347
	4	-.0160	.04186	.995	-.1413	.1093

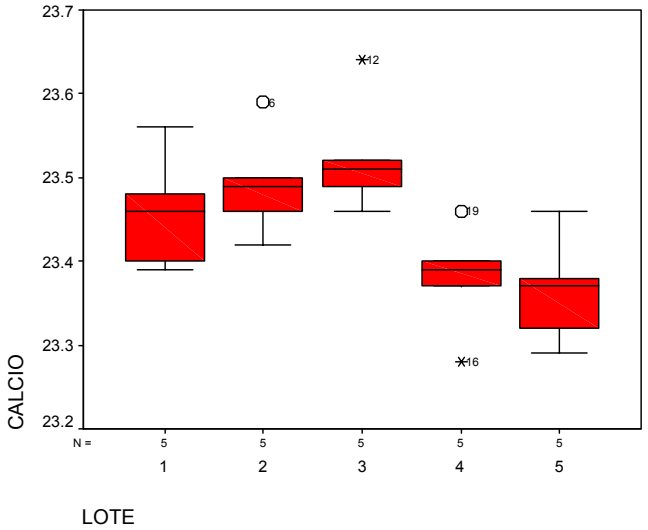
*. The mean difference is significant at the .05 level.

Test of Homogeneity of Variances

CALCIO

Levene Statistic	df1	df2	Sig.
.028	4	20	.998

Resultado de la prueba de homocedasticidad



ANÁLISIS

La tabla del análisis de varianza permite rechazar Ho (Sig. = 0.004, menor a 0.05), es decir existe evidencia de que al menos un par de medias es diferente, surgiendo la pregunta: ¿cuál o cuales son los pares de medias diferentes? Para lo que la comparación de medias de Tukey es la mejor opción para responder a esta interrogante.

La matriz de comparaciones de Tukey compara la media de cada lote con cada uno de los otros lotes, inclusive se señalan los pares de medias que son diferentes con un asterisco.

Esto se puede visualizar mejor en un gráfico boxplot. Donde como ya se mencionó, la diferencia se presenta entre los tratamientos cuyas cajas no se interceptan, aunque esta conclusión se debe reforzar con los valores de la prueba de Tukey.

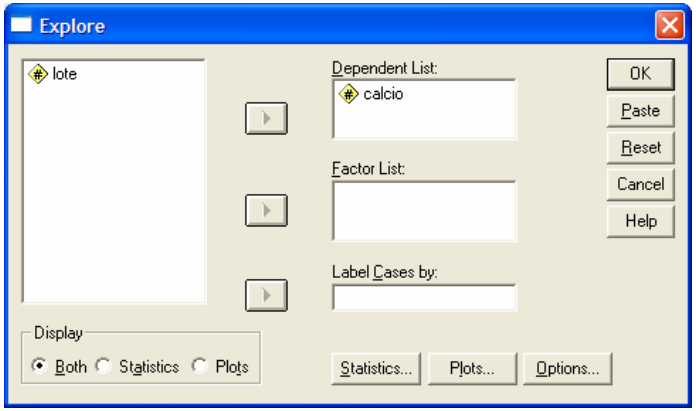
Para darle confiabilidad a las conclusiones se requiere verificar el cumplimiento de supuestos, como la igualdad de varianzas y la normalidad de los residuales. Analizando los valores de la prueba de Levene (sig. = 0.998, muchisimo mayor de 0.05) se observa que no se puede rechazar la hipótesis nula de que todas

las varianzas son estadísticamente iguales, por lo tanto se cumple con la homogeneidad de varianzas.

Otro supuesto a verificar es la normalidad de los datos, el cual se puede revisar con un gráfico de probabilidades normales.

Aunque se observa cierta desviación de la normalidad, ya que debería verse una tendencia lineal. La mejor forma de verificar este supuesto es mediante una prueba numérica, como la **Kolmogorov-Smirnov o Shapiro-Wilks, cuya Ho: es que los datos siguen una distribución normal, contra una Ha: de que los datos no siguen una distribución normal.**

Esto se logra con la secuencia: **ANALYZE -> DESCRIPTIVE STATISTICS -> EXPLORE**, asegurándose de seleccionar la prueba de normalidad en el diálogo **PLOT**.



Nótese que no se define un factor, sólo se trabaja sobre la variable dependiente.

Resultados de Normalidad

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
CALCIO	.134	25	.200*	.978	25	.847

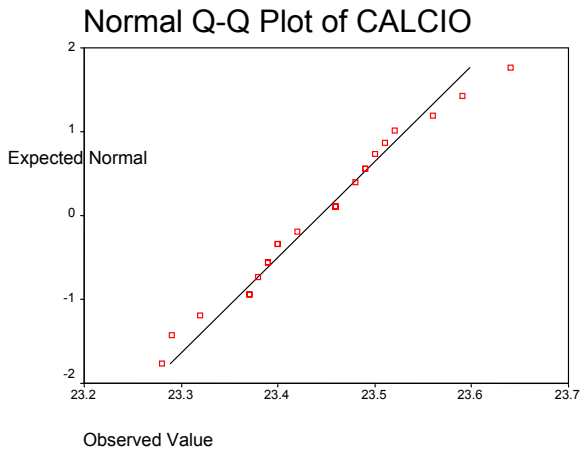
*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Descriptives

			Statistic	Std. Error
CALCIO	Mean		23.4436	.01754
	95% Confidence Interval for Mean	Lower Bound	23.4074	
		Upper Bound	23.4798	
	5% Trimmed Mean		23.4422	
	Median		23.4600	
	Variance		.008	
	Std. Deviation		.08770	
	Minimum		23.28	
	Maximum		23.64	
	Range		.36	
	Interquartile Range		.1100	
	Skewness		.137	.464
	Kurtosis		.034	.902

```
CALCIO Stem-and-Leaf Plot

 Frequency    Stem &  Leaf
    2.00      232 .  89
    1.00      233 .  2
    5.00      233 .  77899
    3.00      234 .  002
    8.00      234 .  66666899
    3.00      235 .  012
    2.00      235 .  69
    1.00 Extremes    (>=23.64)

 Stem width:       .10
 Each leaf:       1 case(s)
```



El valor de **Sig.** indica que no se puede rechazar la Ho y por lo tanto se tiene evidencia de que los datos se comportan como una distribución normal.

Ejemplo 2. Tres diferentes soluciones para lavar están siendo comparadas con el objeto de estudiar su efectividad en el retraso del crecimiento de bacterias en envases de leche de 5 galones. El análisis se realiza en un laboratorio y sólo pueden efectuarse tres pruebas en un mismo día. Se hicieron conteos de colonias durante cuatro días. Analizar los datos y obtener conclusiones acerca de las soluciones.

	Días			
Solución	**1**	**2**	**3**	**4**
I	13	22	18	39
II	16	24	17	44
III	5	4	1	22

Este es un diseño de **Bloques al azar**, donde la variable de bloqueo es días y la variable a comparar es solución. Aquí **ya no se puede realizar** el análisis con el ANOVA ONE-WAY.

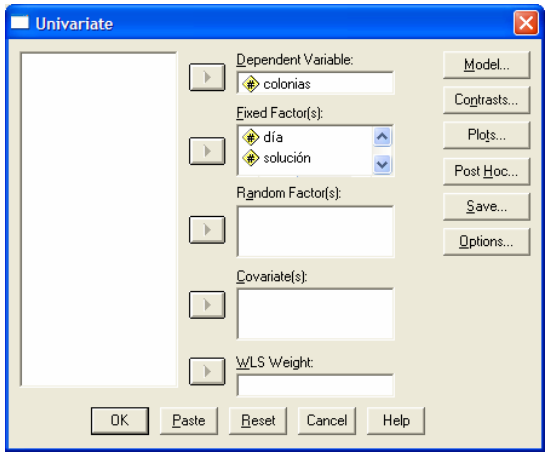
Entonces, la secuencia de análisis es:

1. Capturar los datos en tres columnas, una para identificar la **solución**, otra para el **día** y una para la variable de respuesta **colonias**.

2. Seguir la secuencia

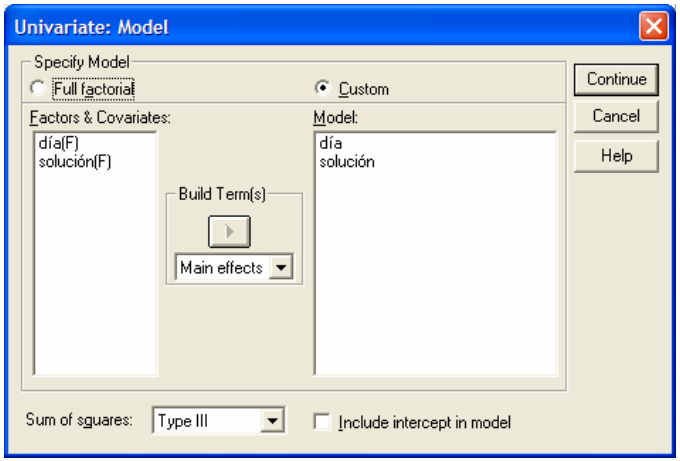
ANALYZE -> GENERAL LINEAR MODEL -> UNIVARIATE

3. Colocar en la caja de diálogo la variable dependiente y los factores fijos, en este caso, día y solución

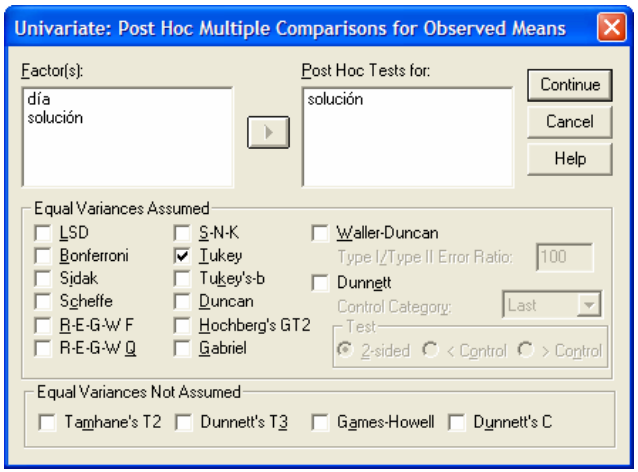


4. Abrir el diálogo **Model** y seleccionar en **specify model** la opción **custom.**

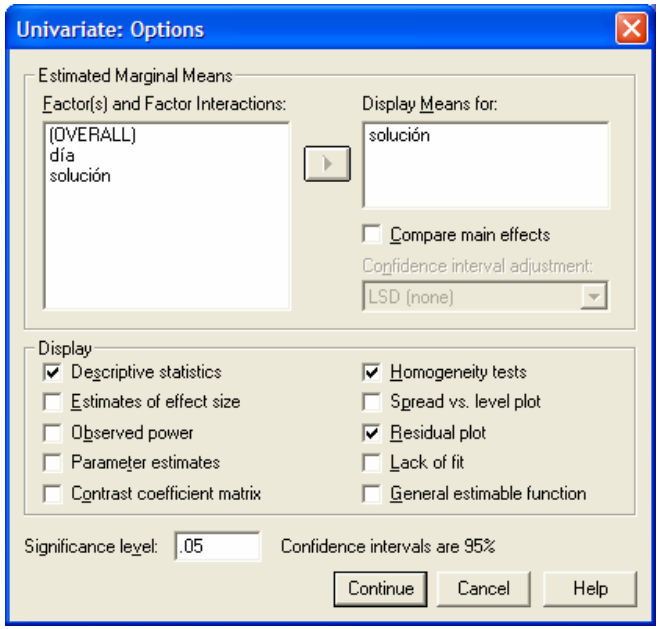
En la caja **model** ingresar las **variables día y solución**. Además asegurarse de que en **Build Term(s)** este la opción **Main effects**, que **Sum squares** esté en **Type III** y que **no esté activa** la opción **Include intercept in model**



5. Abrir el diálogo Post Hoc, y ahí seleccionar la prueba de comparación de medias a realizar. Recuerden que la variable de interés es solución.



6. Abrir el diálogo OPTIONS y seleccionar las opciones a trabajar.



7. Ahora si, listos para realizar el análisis.

RESULTADOS

Univariate Analysis of Variance

Between-Subjects Factors

		N
DÍA	1	3
	2	3
	3	3
	4	3
SOLUCIÓN	1	4
	2	4
	3	4

Lista de los factores en estudio

Tests of Between-Subjects Effects

Dependent Variable: COLONIAS

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	6029.167[a]	6	1004.861	116.318	.000
DÍA	1106.917	3	368.972	42.711	.000
SOLUCIÓN	703.500	2	351.750	40.717	.000
Error	51.833	6	8.639		
Total	6081.000	12			

a. R Squared = .991 (Adjusted R Squared = .983)

Resultados de ANVA, sólo es de interés la F y la Sig. de solución.

Estimated Marginal Means

SOLUCIÓN

Dependent Variable: COLONIAS

SOLUCIÓN	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	23.000	1.470	19.404	26.596
2	25.250	1.470	21.654	28.846
3	8.000	1.470	4.404	11.596

Intervalos de confianza para cada tratamiento.

Post Hoc Tests

Los valores de P indican que hay evidencia estadística de diferencias entre las soluciones, ahora hay que decir cuales son las que realmente son diferentes y cuál seria la mejor. Para esto hay que realizar una prueba de Tukey, tomando el

Multiple Comparisons

Dependent Variable: COLONIAS
Tukey HSD

(I) SOLUCIÓN	(J) SOLUCIÓN	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-2.2500	2.07833	.558	-8.6269	4.1269
	3	15.0000*	2.07833	.001	8.6231	21.3769
2	1	2.2500	2.07833	.558	-4.1269	8.6269
	3	17.2500*	2.07833	.000	10.8731	23.6269
3	1	-15.0000*	2.07833	.001	-21.3769	-8.6231
	2	-17.2500*	2.07833	.000	-23.6269	-10.8731

Based on observed means.

*. The mean difference is significant at the .05 level.

valor del CME de la tabla de ANVA.

Después hay que verificar los supuestos del Análisis de Varianza, mediante las secuencias de análisis ya conocidas.

ONEWAY ANOVA

Test of Homogeneity of Variances

COLONIAS

Levene Statistic	df1	df2	Sig.
.144	2	9	.868

El valor de Sig=0.868 nos indica que se cumple la homocedasticidad

Explore

La significancia de la prueba de normalidad indica que los datos cumplen con este supuesto, ya que no se rechaza Ho.

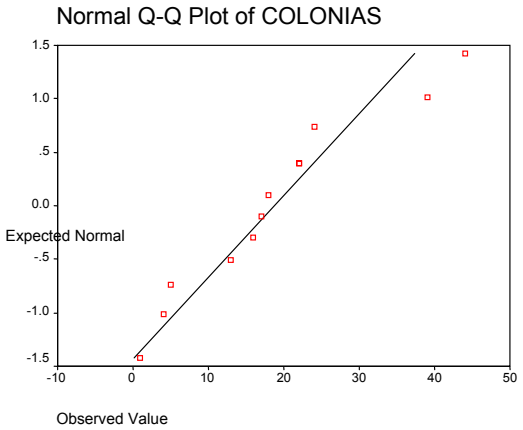
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
COLONIAS	.177	12	.200*	.932	12	.397

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

COLONIAS



El gráfico no muestra desviaciones severas de la normalidad, por lo que se cumplen los supuestos del Análisis de Varianza y las conclusiones son totalmente válidas y confiables

DISEÑOS FACTORIALES

Este tipo de diseños permiten analizar varios factores a la vez, considerando su interacción.

La construcción típica de un factorial **axb** se presenta a continuación, donde **a** indica el número de niveles del primer factor y **b** el del segundo factor.

	FACTOR B				TOTAL
FACTOR A	1	2	. . .	b	$Y_{i..}$
1	$Y_{111}, Y_{112}, Y_{113}, Y_{114}$	$Y_{121}, Y_{122}, Y_{123}, Y_{124}$	. . .	$Y_{1b1}, Y_{1b2}, Y_{1b3}, Y_{1b4}$	$Y_{1..}$
2	$Y_{211}, Y_{212}, Y_{213}, Y_{214}$	$Y_{221}, Y_{222}, Y_{223}, Y_{224}$	. . .	$Y_{1b1}, Y_{1b2}, Y_{1b3}, Y_{1b4}$	$Y_{2..}$
. . .	. . .	. . .	. . .	. . .	. . .
A	$Y_{a11}, Y_{a12}, Y_{a13}, Y_{a14}$	$Y_{a21}, Y_{a22}, Y_{a23}, Y_{a24}$	. . .	$Y_{ab1}, Y_{ab2}, Y_{ab3}, Y_{ab4}$	$Y_{a..}$
Total $Y_{.j.}$	$Y_{.1.}$	$Y_{.2.}$	...	$Y_{.b.}$	Y...

Modelo: $Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk}$
con: i = 1,2,3, ... a; j = 1,2,3, ... b; k = 1,2,3, ..., k

1. **Ho:** $\tau_i = 0$ v.s. **Ha:** $\tau_i \neq 0$; para al menos una i.
2. **Ho:** $\beta_j = 0$ v.s. **Ha:** $\beta_j \neq 0$; para al menos una j.
3. **Ho:** $\tau_i\beta_j = 0$ v.s. **Ha:** $\tau_i\beta_j \neq 0$ para al menos un par $i \neq j$.

$$SC_{TOTAL} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \overline{Y}_{...})^2$$

$$SC_A = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{i..} - \overline{Y}_{...})^2 = bn\sum_{i=1}^{a}(Y_{i..} - \overline{Y}_{...})^2$$

$$SC_B = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{.j.} - \overline{Y}_{...})^2 = an\sum_{j=1}^{b}(Y_{.j.} - \overline{Y}_{...})^2$$

$$SC_{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...})^2 =$$

$$= n\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...})^2$$

$$SC_{ERROR} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \overline{Y}_{ij.})^2$$

GRADOS DE LIBERTAD

A = a - 1
B = b - 1
AB = (a - 1)(b - 1)
ERROR = ab(n - 1)
TOTAL = abn - 1

¿TABLA DE ANALISIS DE VARIANZA, PARA UN DISEÑO: **A**x**B**x**C**x**D**?

Se puede ver a través de algunos ejemplos

Ejemplo 3. Se encuentra en estudio el rendimiento de un proceso químico. Se cree que las dos variables más importantes son la temperatura y la presión. Seleccionando para el estudio tres temperaturas y tres presiones diferentes, obteniendo los siguientes resultados de rendimiento.

Temperatura\Presión	Baja	Media	Alta
Baja	90.4	90.7	90.2
	90.2	90.6	90.4
Intermedia	90.1	90.5	89.9
	90.3	90.6	90.1
Alta	90.5	90.8	90.4
	90.7	90.9	90.1

Diseño factorial 3x3 con 2 repeticiones.

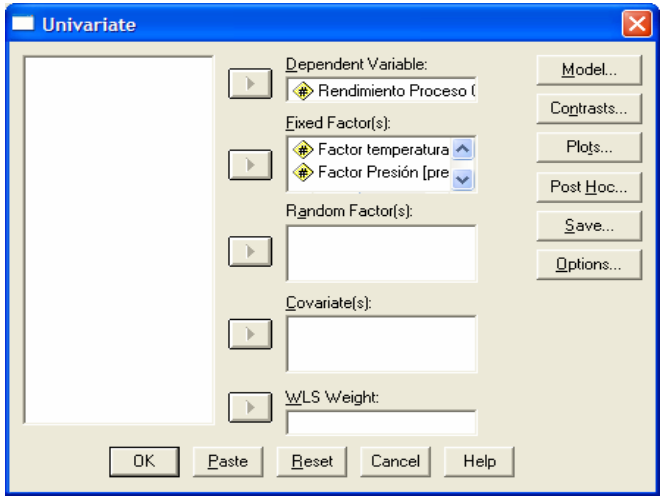
El análisis requiere los siguientes pasos:

1. Capturar los datos en tres columnas, una para identificar el **factor temperatura**, otra para **presión** y una para la variable de respuesta **rendimiento,** con 18 datos o renglones.

2. Seguir la secuencia

ANALYZE -> GENERAL LINEAR MODEL -> UNIVARIATE

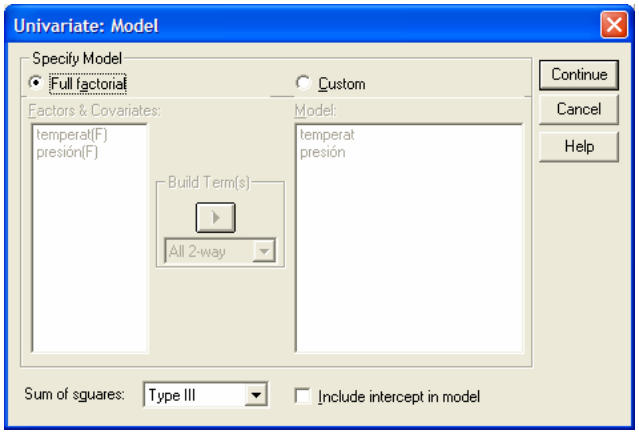
3. Colocar en la caja de diálogo la variable dependiente y los factores fijos, en este caso, temperatura y presión



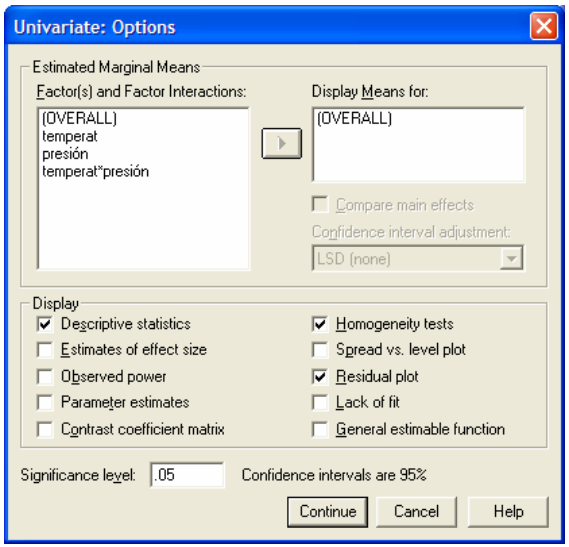
4. Abrir el diálogo **Model** y seleccionar en **specify model** la opción **Full factorial.**

En la caja **model** ingresar los **factores temperatura y presión**. Además asegurarse de que **Sum squares** esté en **Type III** y que **no esté activa** la opción **Include intercept in model**.

5. Cuidar que en el diálogo Post Hoc, NO se seleccione ninguna opción, ya que en un diseño factorial son de mayor interés los efectos de interacción y no tanto los efectos principales.

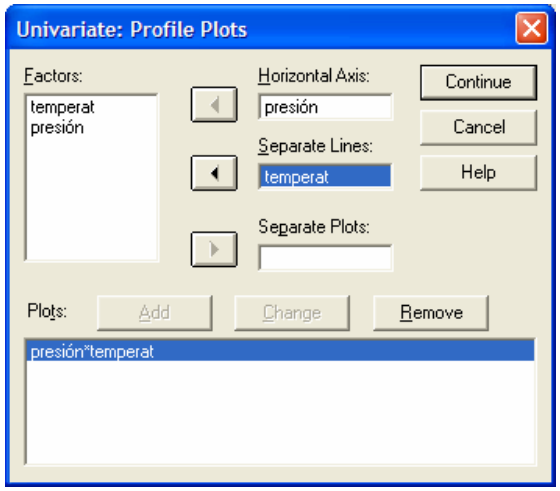


6. Abrir el diálogo OPTIONS y seleccionar las opciones a trabajar.



7. Abrir el diálogo PLOTS y

Seleccionar un factor para el **eje horizontal** y e otro factor para **líneas separadas**



8. Ahora si, listos para realizar el análisis.

RESULTADOS

Between-Subjects Factors

		Value Label	N
Factor temperatura	1	Baja	6
	2	Media	6
	3	Alta	6
Factor Presión	1	Baja	6
	2	Media	6
	3	Alta	6

Descriptive Statistics

Dependent Variable: Rendimiento Proceso Químico

Factor temperatura	Factor Presión	Mean	Std. Deviation	N
Baja	Baja	90.3000	.14142	2
	Media	90.6500	.07071	2
	Alta	90.3000	.14142	2
	Total	90.4167	.20412	6
Media	Baja	90.2000	.14142	2
	Media	90.5500	.07071	2
	Alta	90.0000	.14142	2
	Total	90.2500	.26646	6
Alta	Baja	90.6000	.14142	2
	Media	90.8500	.07071	2
	Alta	90.2500	.21213	2
	Total	90.5667	.29439	6
Total	Baja	90.3667	.21602	6
	Media	90.6833	.14720	6
	Alta	90.1833	.19408	6
	Total	90.4111	.27630	18

Tabla de medias que permite elaborar los gráficos de interacciones

Tests of Between-Subjects Effects

Dependent Variable: Rendimiento Proceso Químico

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	147136.180[a]	9	16348.464	919601.1	.000
TEMPERAT	.301	2	.151	8.469	.009
PRESIÓN	.768	2	.384	21.594	.000
TEMPERAT * PRESIÓN	6.889E-02	4	1.722E-02	.969	.470
Error	.160	9	1.778E-02		
Total	147136.340	18			

a. R Squared = 1.000 (Adjusted R Squared = 1.000)

Resultados de ANVA, donde son de interés las F's y las Sig.'s de los efectos principales y los de la interacción.

ANÁLISIS

En la tabla de Análisis de Varianza, se observan efectos significativos para:

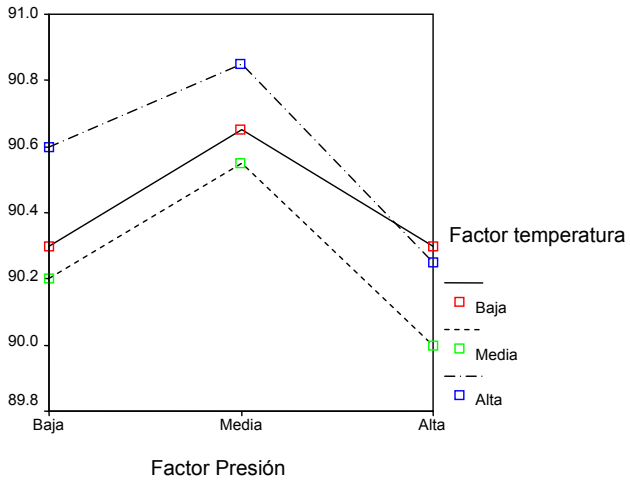
Modelo, al menos uno de los parámetros del modelo son diferentes de cero.

Temperatura, Si hay efecto de temperatura (significancia 0.009)

Presión, SI hay efecto de Presión (significancia 0.000)

Interacción T*P, NO hay efecto conjunto de los dos factores.

Pero entonces, cual es la combinación de factores que genera mayor rendimiento. Esto se ve en el gráfico de interacciones, donde se aprecia que la combinación: Presión media y temperatura alta se tiene el mayor rendimiento.



EJERCICIOS

1. Se encuentra bajo estudio el efecto que tienen 5 reactivos distintos (A, B, C, D y E) sobre el tiempo de reacción de un proceso químico. Cada lote de material nuevo es lo suficientemente grande para permitir que sólo se realicen 5 ensayos. Más aún, cada ensayo tarda aproximadamente una hora y media por lo que sólo pueden realizarse cinco ensayos por día. En el experimento se busca controlar sistemáticamente las variables lote de material y día, ¿que se puede decir del tiempo de reacción de los 5 reactivos diferentes?

	Día				
Lote	**1**	**2**	**3**	**4**	**5**
1	A,8	B,7	D,1	C,7	E,3
2	C,11	E,2	A,7	D,3	B,8
3	B,4	A,9	C,10	E,6	D,5
4	D,6	C,8	E,6	B,1	A,10
5	E,4	D,2	B,3	A,8	C,8

Recomendación: Diseño de Cuadrados Latinos, con día y lote como variables de bloqueo. Se hace igual que el ejemplo 2, pero en el modelo también se incluye la variable lote.

2. En un experimento para comparar el porcentaje de eficiencia de cuatro diferentes resinas quelantes (A, B, C y D) en la extracción de iones de Cu^{2+} de solución acuosa, el experimentador sólo puede realizar cuatro corridas con cada resina. De manera que durante tres días seguidos se preparo una solución fresca de iones Cu^{2+} y se realizó la extracción con cada una de las resinas, tomadas de manera aleatoria, obteniendo los siguientes resultados. ¿Cuál es el modelo más adecuado para analizar este experimento y cuales son sus conclusiones?

Día	A	B	C	D
1	97	93	96	92
2	90	92	95	90
3	96	91	93	91
4	95	93	94	90

Recomendación: Diseño de Bloques al azar, con día como variable de bloqueo.

3. Se llevó a cabo un experimento para probar los efectos de un fertilizante nitrogenado en la producción de lechuga. Se aplicaron cinco dosis diferentes de nitrato de amonio a cuatro parcelas (réplicas). Los datos son el número de lechugas cosechadas de la parcela

Tratamiento Kg N/Ha				
0	104	114	90	140
50	134	130	144	174
100	146	142	152	156
150	147	160	160	163
200	131	148	154	163

Recomendación: Diseño Completamente al Azar (One-Way)

4.. En una operación de lotes se produce un químico viscoso, donde cada lote produce suficiente producto para llenar 100 contenedores. El ensayo del producto es determinado por análisis infrarrojo que realiza duplicado alguno de los 20 analistas del laboratorio. En un esfuerzo por mejorar la calidad del producto se realizó un estudio para determinar cual de tres posibles fuentes de variabilidad eran significativas en el proceso y su magnitud.

Las fuentes seleccionadas fueron: la variable **A** lotes, se seleccionaron aleatoriamente tres lotes de producción mensual, la variable analistas, **B**, seleccionando dos de manera aleatoria, la variable **C** corresponde a dos contenedores seleccionados de manera aleatoria de cada lote.

Lote	No. de Contenedor			
	I		II	
	Analista		Analista	
	M	P	M	P
23	94.6	95.8	97.7	97.8
	95.2	95.8	98.1	98.6
35	96.2	96.5	98.0	99.0
	96.4	96.9	98.4	99.0
2	97.9	98.4	99.2	99.6
	98.1	98.6	99.4	100.0

Recomendación: Diseño factorial 3x2x2, lo que interesa es la significancia de los efectos principales, así como la de cada una de las dobles interacciones y la triple interacción.

Capítulo 5

Análisis de Regresión

Problemas que se plantean:

1) ¿Cuál es el modelo matemático más apropiado para describir la relación entre una o más variables independientes (X´s) y una variable dependiente (Y)?

2) Dado un modelo especifico, ¿qué significa éste y cómo se encuentran los parámetros del modelo que mejor ajustan a nuestros datos? Si el modelo es una línea recta: ¿cómo se encuentra la "mejor recta"?

La ecuación de una línea recta es:

$$Y = f(x) = \beta_0 + \beta_1 X$$

β_0 ordenada al origen β_1 pendiente

En un análisis de regresión lineal simple, el problema es encontrar los valores que mejor estimen a los parámetros β_0 y β_1. A partir de una muestra aleatoria.

El modelo de regresión lineal es:

$$Y_i = \mu_{y/Xi} + \varepsilon_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (i = 1,2,3, ..., n)$$

Para cada observación el modelo es:

$Y_1 = \beta_0 + \beta_1 X_1 + \varepsilon_1$

$Y_2 = \beta_0 + \beta_1 X_2 + \varepsilon_2$

. . .

$Y_n = \beta_0 + \beta_1 X_n + \varepsilon_n$

El cual se puede escribir como:

$$_n\mathbf{y}_1 = \begin{pmatrix} Y_1 \\ Y_2 \\ . \\ Y_n \end{pmatrix} \qquad _n\mathbf{X}_2 = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ . & . \\ . & . \\ 1 & X_n \end{pmatrix} \qquad _2\boldsymbol{\beta}_1 = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad _n\boldsymbol{\varepsilon}_1 = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{pmatrix}$$

donde:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ . & . \\ . & . \\ . & . \\ 1 & X_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \\ \\ \\ \varepsilon_n \end{pmatrix} = \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Estimación por mínimos cuadrados

Sea $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ la respuesta estimada en X_i con base en la línea de regresión ajustada. La distancia vertical entre el punto (X_i, Y_i) y el punto (X_i, $\hat{Y}_i$) de la recta ajustada está dada por el valor absoluto de $|Y_i - \hat{Y}_i|$ o $|Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)|$, cuya suma de cuadrados es:

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

El problema ahora es encontrar los valores de β_0 y β_1 ($\hat{\beta}_0$ y $\hat{\beta}_1$) tales que $\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$ sea mínimo.

Solución:

Si $Q = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$, entonces

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \qquad (1)$$

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)\ (-X_i) = 0 \qquad (2)$$

que conduce a las Ecuaciones Normales de Mínimos Cuadrados

$$\Sigma Y_i - \Sigma\beta_0 - \Sigma\beta_1 X_i = \Sigma Y_i - n\beta_0 - \beta_1 \Sigma X_i \qquad .. (1')$$

$$-\Sigma X_i Y_i + \beta_0 \Sigma X_i + \beta_1 \Sigma X_i X_i \qquad .. (2')$$

ordenando

$\beta_0 n + \beta_1 \Sigma X_i = \Sigma Y_i$
$\beta_0 \Sigma Xi + \beta_1 \Sigma X_i^2 = \Sigma X_i Y_i$

$$\begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix}\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \sum Y_i \\ \sum X_i Y_i \end{pmatrix}$$

X'Xβ = X'y

β = (X'X)$^{-1}$X'y

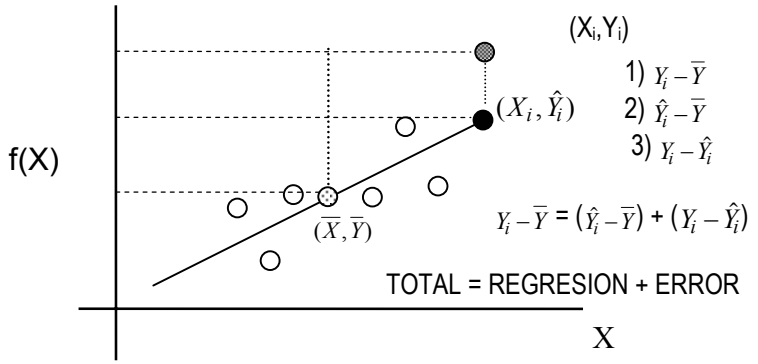
Solución matricial para para calcular los parámetros de la ecuación de regresión

La solución algebraica de las ecuaciones normales, para datos muestrales, genera las siguientes ecuaciones:

$$b_0 = \frac{\sum_{i=}^{n} Y_i \sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}$$

$$b_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}$$

ALGO DE GEOMETRÍA



Al aplicar sumatorias y elevar al cuadrado se tiene:

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}[(\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i)]^2$$
$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + \sum(Y_i - \hat{Y}_i)^2$$

S.C.$_{TOTAL}$ = S.C.$_{REGRESION}$ + S.C.$_{ERROR}$

Cantidades que permiten realizar un ANVA, para contrastar las hipótesis:
Ho: $\beta_i = 0$ v.s. **Ha**: $\beta_i \neq 0$.

F.V.	g.l.	S.C.	C.M.	Fc	Ft
REGRESION	1			$\frac{CM_{REGRESION}}{CM_{RESIDUAL}}$	$F_{1,n-2,\alpha}$
RESIDUAL	N - 2				
TOTAL	N - 1				

Este ANVA tiene el siguiente juego de hipótesis:

Ho: β_i=0, es decir que todos los coeficientes del modelo son iguales a cero y por lo tanto no hay un modelo lineal que describa el comportamiento de los datos.

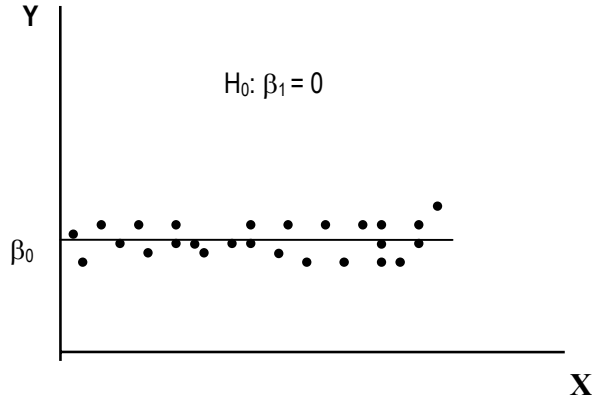
Contra **Ha:** de que al menos uno de los coeficientes es diferente de cero y entonces si hay un modelo lineal.

INTERPRETANDO a β_0 y β_1

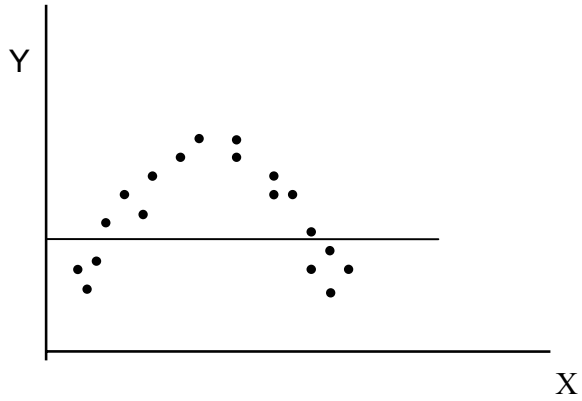
H_0: β_1 = 0

Caso 1.- **H_0: β_1 = 0 No se rechaza. Es decir que la pendiente es cero o que no hay pendiente,** entonces se tienen dos opciones de interpretación.

a) Si la suposición de línea recta es correcta significa que X no proporciona ayuda para predecir Y, esto quiere decir que $\overline{Y}$ predice a Y.



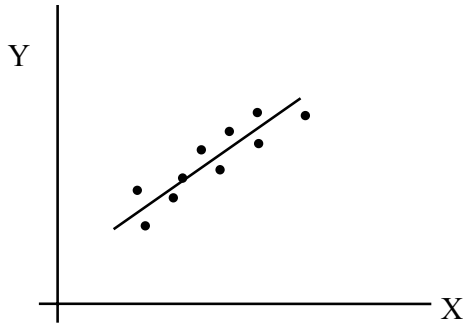
b) La verdadera relación entre X e Y no es lineal, esto significa que el modelo puede involucrar funciones cuadráticas cúbicas o funciones más complejas.



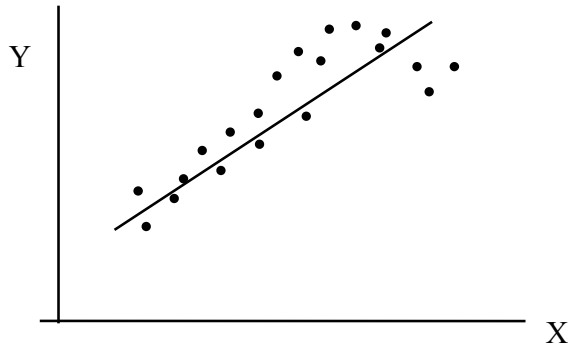
NOTA: Si hay una curvatura se requiere un elemento cuadrático en el modelo, si hay dos curvaturas entonces se requiere un cúbico y así sucesivamente.

Caso 2.- Ho: $\beta_1 = 0$ SI se rechaza (es decir, si hay pendiente o en otras palabras si hay un modelo lineal que describe el comportamiento de los datos).

a) X proporciona información significativa para predecir Y



b). El modelo puede tener un término lineal más, quizás un término cuadrático.



Caso 3. Prueba. Ho: $\beta_0 = 0$

Si NO se rechaza esta Hipótesis, puede ser apropiado ajustar un modelo sin β_0, siempre y cuando exista experiencia previa o teoría que sugiera que la recta ajustada debe pasar por el origen y que existan datos alrededor del origen para mejorar la información sobre β_0.

CORRELACION

Si X e Y son dos variables aleatorias, entonces el coeficiente de correlación se define como:

1) $r \in [-1,1]$

2) r es independiente de las unidades de X e Y

3) $\hat{\beta}_1 > 0 \Leftrightarrow r > 0$

$\hat{\beta}_1 < 0 \Leftrightarrow r < 0$

$\hat{\beta}_1 = 0 \Leftrightarrow r = 0$

r es una medida de la fuerza de asociación lineal entre X e Y

NOTA: NO se puede ni se deben establecer relaciones causales a partir de los valores de r, ya que ambas variables son aleatorias.

COEFICIENTE DE DETERMINACIÓN r^2

$$r^2 = \frac{SC_{total} - SC_{error}}{SC_{total}} = \frac{SC_{regresión}}{SC_{total}}$$

donde $r^2 \in [0,1]$

Esta r-cuadrada es una medida de la variación de Y explicada por los cambios o variación en la X. Es común leerla como porcentaje de variación en Y explicada por los cambios en X.

REGRESION NO-LINEAL

En ocasiones, la relación X-Y presenta una tendencia curvilínea, entonces se debe recurrir a los ajustes no lineales. En estos casos es importante tener una idea más o menos clara del tipo de curva al que se debe ajustar, ya que hay: logarítmicas, cuadráticas, cúbicas, inversas, potenciales, logísticas o exponenciales, entre otras opciones.

REGRESION LINEAL MULTIPLE

Esta regresión se refiere a modelos lineales cuando se consideran dos o más variables independientes.

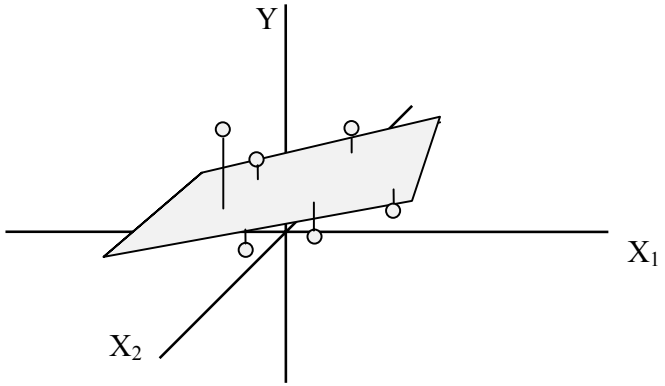
$$Y = f(X_1, X_2, ..., X_K) = f(\mathbf{x})$$

Comparando la regresión simple contra la múltiple se tiene que:

1) Es más difícil la elección del mejor modelo, ya que casi siempre hay varias opciones razonables.

2) Se dificulta visualizar el modelo, por la dificultad de "pintar" más de tres dimensiones.

3) Requiere cálculos complejos, generalmente se realiza con recursos computacionales y software especializado.



Ajuste de un plano lineal con dos variables independientes.

Mínimos Cuadrados. Al igual que en la regresión lineal simple, se puede trabajar el método de mínimos cuadrados**.** Para esto:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

donde:

$\varepsilon = Y_i - (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k)$

En base a los datos muestrales

$$Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_o + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k)$$

Al elemento de la derecha se le conoce como residual y refleja la desviación de los datos observados con respecto a la línea o plano ajustado.

Suma de cuadrados, elevando al cuadrado y sumando los elementos de la ecuación anterior.

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k))^2$$

El método consiste en encontrar los valores $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots$ llamados estimadores de mínimos cuadrados, para los cuales la suma de cuadrados es mínima. De tal manera que, se pueda construir la siguiente tabla de ANVA.

Tabla de ANVA para la hipótesis Ho: $\beta_i = 0$; Ha: al menos un $\beta_i \neq 0$

F.V.	g.l.	S.C.	C.M.	F	r^2
Regresión	k	$SC_{Tot} - SC_{res}$	SC_{Reg}/k	$CM_{Reg}/CMres$	$(SC_{Tot}-SC_{error})/SC_{Tot}$
Residual o Error	n-k-1	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$SC_{res}/(n-k-1)$		
Total	n-1	$\sum_{i=1}^{n}(Y_i - \overline{Y})^2$			

Donde los supuestos del análisis de Regresión se pueden resumir en la siguiente expresión.

$$\varepsilon \sim NI(\mu_{Y/X1,X2,\ldots,Xk}, \sigma^2)$$

Los errores o residuales se distribuyen normal e independientemente con desviaciones al ajuste lineal igual a cero y varianza σ^2.

CORRELACIÓN PARCIAL y PARCIAL MÚLTIPLE

Medida de la fuerza de relación lineal entre dos variables, después de controlar los efectos de otras variables en el modelo.
Cuya representación está dada por:

$R_{y,x1/x2}$ $R_{y,x1/x2,x3}$ $R_{y,(x3,x4,x5)/x1,x2}$

Expresiones que se leen:

$R_{y,x1/x2}$ Correlación de las variables Y-X_1, cuando se tiene controlado el efecto de X_2 en un modelo. También se puede leer: correlación de Y-X_1, cuando X_2 ya está en el modelo.

$R_{y,x1/x2,x3}$ Correlación de las variables Y-X_1, cuando se tienen controlados los efectos de X_2 y X_3 en un modelo.

$R_{y,(x3,x4,x5)/x1,x2}$ Correlación de las variables X_3, X_4 y X_5 con Y, cuando se tienen controlados los efectos de X_1 y X_2 en un modelo.

CORRELACION Y DETERMINACIÓN MÚLTIPLE

$$R_{yx_1,x_2,\ldots,x_k} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(\hat{Y}_i - \overline{\hat{Y}})}{\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \sum_{i=1}^{n}(\hat{Y}_i - \overline{\hat{Y}})^2}}$$

$$R^2_{y/x1,x2,\ldots,xk} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2 - \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = (SC_{total} - SC_{error})/SC_{total}$$

Donde r y r^2 representan la correlación y determinación simple, mientras que R y R^2 se utilizan para la correlación y determinación múltiple.

F's PARCIALES

La F's parciales son una herramienta útil para verificar si el ingreso o eliminación de una variable o grupos de variable mejoran el ajuste de un modelo lineal.

Este verificación se inicia con algunas preguntas, suponiendo 3 variables X_1, X_2 y X_3

1) ¿Se puede predecir el valor de Y utilizando sólo X_1?

2) ¿Adicionar X_2 contribuye significativamente en la predicción de Y, una vez que se considera la contribución de X_1?

3) ¿Contribuye X_3, dados X_1 y X_2 en el modelo?

Las respuestas a estas preguntas se obtienen al contrastar las siguientes hipótesis:

Ho: La adición de X* al modelo, incluyendo X_1, X_2, ..., X_k, no mejora significativamente la predicción de Y.

Ho: $\beta^* = 0$, donde β^* es el coeficiente de X*, en la ecuación de regresión.
Ha: $\beta^* \neq 0$

Cuyo estadístico de prueba es:

$$t_o = \frac{\hat{\beta}^*}{S_{\hat{\beta}^*}}$$

Cuya regla de decisión es: rechazar Ho si $t_o > t_{\alpha/2,n-k-1}$

ASPECTOS PRÁCTICOS DE LA REGRESIÓN LINEAL MÚLTIPLE

Para determinar la relación entre dos o más variables de regresión X y la respuesta Y. El problema general consiste en ajustar el modelo.

$$y= \beta o +\beta_1 X_1+\beta_2 X_2+...+\beta_k X_k+\varepsilon$$

Usualmente los parámetros desconocidos (β_k) se denominan coeficientes de regresión y pueden determinarse mediante **mínimos cuadrados**. Donde ε denominado error aleatorio debe presentar una media igual a cero y su varianza σ^2 no debe estar correlacionadas.

Pruebas de hipótesis de la regresión lineal múltiple

A menudo se desea probar que tan significantes son los parámetros del modelo de regresión, lo cual se logra al contrastar si dichos coeficientes son iguales a cero; las hipótesis son:

$$Ho: \beta_o=\beta_1=..=\beta_k=0$$

$$Ha: \beta_i \neq 0$$

Rechazar Ho implica que al menos una de variables del modelo contribuye significativamente al ajuste. El parámetro para probar esta hipótesis es una generalización del utilizado en regresión lineal simple. La suma total de cuadrados (SC_y) se descompone en la suma de cuadrados de regresión (SC_R) y en la sumas de cuadrados del error (SC_e).

$$SC_y = SC_R + -SC_e$$

Consecuentemente el valor de F estimado se obtiene de la ecuación:

$$F_o = \frac{SC_R / k}{SC\varepsilon / (n-k-1)} = \frac{CM_R}{CM_\varepsilon}$$

Valor que se compara con una $F_{\alpha,\text{g.l. numerador(k), g.l. denominador(n-k-1)}}$ de tablas. La regla de decisión es: Rechazar Ho si F_o > F de tablas.

Criterio para la selección de variables

Es importante probar las hipótesis con respecto a los coeficientes de regresión individuales; tales pruebas son útiles para evaluar cada variable de regresión en el modelo. En ocasiones el modelo puede ser más efectivo si se le introducen variables adicionales o, quizá si se desechan una o más variables que se encuentran en el mismo.

Introducir variables al modelo de regresión provoca que la suma de cuadrados de la regresión aumente y que la del error disminuya. Se debe decidir si el incremento de la suma de cuadrados de la regresión es suficiente para garantizar el uso de la variable adicional en el modelo. Además si se agrega una variable poco importante al modelo se puede aumentar el cuadrado medio del error, disminuyendo así la utilidad del mismo.

La hipótesis para probar la significancia de cualquier coeficiente individual, por ejemplo β_i son:

$$\text{Ho: } \beta_i = 0$$
$$\text{Ha: } \beta_i \neq 0$$

Y la estadística apropiada para probar la ecuación es:

$$t_o = \frac{\beta_i}{s_{B_i}}$$

Donde β_i es el coeficiente a contrastar y $s_{\beta i}$ es el error estándar del coeficiente a contrastar. La regla de decisión es: rechazar Ho si $|t_o| > t_{\alpha/2,n-k-1}$

Coeficiente de determinación R^2 y R^2 ajustado

Después de encontrar la recta de regresión, se debe de investigar que tan bien se ajusta a los datos mediante el calculo de R^2.

Este coeficiente se construye con base en dos cantidades. La primera es la suma de los cuadrados minimizada denominada suma de cuadrados del error (SC_E), la cual representa la suma de las desviaciones al cuadrado de los datos a la recta que mejor se ajusta. La segunda cantidad es la suma de cuadrados alrededor de la media $\bar{Y}$, y se conoce como la suma de cuadrados totales ($SC_{Tot.}$).

El valor de R^2 se define de la siguiente forma:

$$R^2 = \frac{SS_{Tot.} - SS_E}{SS_{Tot.}} = \frac{SS_R}{SS_{Tot.}} = 1 - \frac{SS_E}{SS_{Tot.}}$$

Y se interpreta como el porcentaje de la suma de cuadrados total que es explicada por la relación lineal. Conviene aclarar que a pesar que R^2 es un buen indicador de la calidad del ajuste de regresión, no se debe usar como un criterio único de selección del modelo.

Al agregar variables, a un modelo lineal, el coeficiente de correlación y de determinación siempre aumentan. Por lo que es importante no tomar como único criterio de selección de modelos el valor de R o R^2. Es mejor utilizar el coeficiente de determinación ajustado, que considera el número de variables independientes (X's) en el modelo, y cuya fórmula de cálculo es.

$$R^2_{ajustada} = R^2_a = 1 - \frac{\frac{SC_{Error}}{n-k}}{\frac{SC_{Tot}}{n-1}} = 1 - \frac{(n-1)}{(n-k)}\frac{SC_{Error}}{SC_{Tot.}}$$

El criterio de selección de variables para un ajuste lineal es que la R^2 sea mayor y que el cuadrado medio del error sea más pequeño. De tal manera que al comparara dos o más modelos el mejor es aquel con R^2 mayor y menor CM_{Error} .

MÉTODOS DE SELECCIÓN DE VARIABLES

Para realizar un ajuste lineal múltiple existen se tienen tres métodos clásicos de selección de variables.

- FORWARD, implica ir "metiendo" variables al modelo en función de su significancia, evitando que entren las no significativas.
- BACKWARD "mete" todas las variables al modelo y empieza a sacar las menos significativas hasta quedarse únicamente con las significativas.
- STEPWISE combina los dos métodos anteriores para "meter" y "sacar" variables hasta quedarse con las más significativas.

DESPUÉS DEL ANÁLISIS DE REGRESIÓN

Hay que verificar supuestos, así como cuidar problemas de multicolinealidad y autocorrelación. También se puede realizar una prueba de falta de ajuste, la cual contrasta la Ho: de que el modelo se ajusta y describe los datos muestrales (requiere tener repeticiones de Y para cada uno de los valores de X).

EJEMPLOS

Ejemplo 1. Relación de gastos médicos mensuales en relación con el tamaño de familia **(REGRESIÓN LINEAL SIMPLE).**

TAMAÑO DE FAMILIA	GASTOS MEDICOS MENSUALES (en dólares)
2	20
2	28
4	52
5	50
7	78
3	35
8	102
10	88
5	51
2	22
3	29
5	49
2	25

¿Existe evidencia para establecer una relación lineal entre el tamaño de la familia y los gastos médicos? ¿Si la respuesta es afirmativa, cual es la ecuación de esta relación?

¿Se cumplen los supuestos del análisis de regresión?

Secuencia de análisis

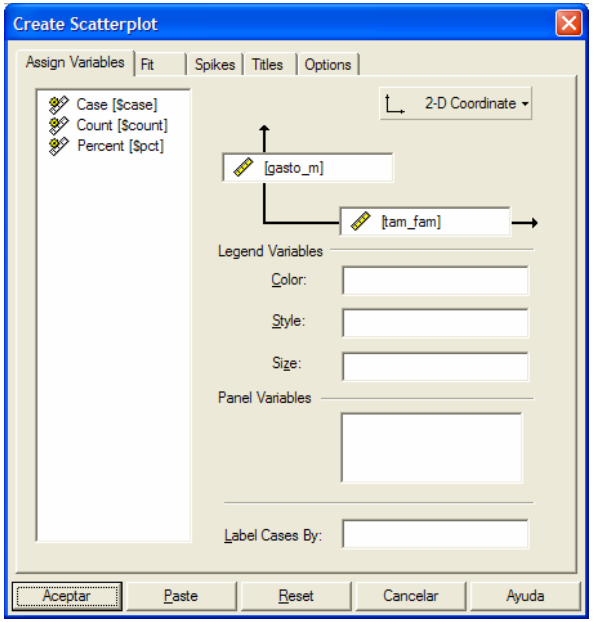
1. Crear el archivo de datos con dos columnas, una para la variable independiente (X) y otra para la variable dependiente (Y).

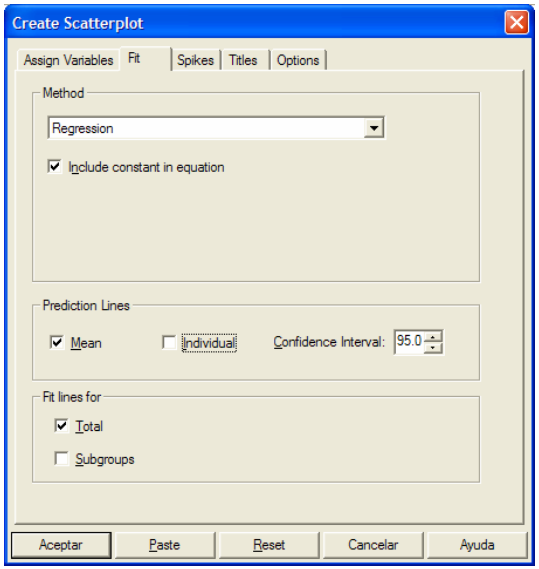
2. El primer paso es realizar un diagrama de dispersión que muestre la tendencia de los datos. Esto se hace con la opción del menú:
   GRAPH -> INTERACTIVE -> SCATTERPLOT

3. Colocar las variables en el eje correspondiente

Notar las opciones tipo fólder que se presentan en esta caja de diálogo. Seleccionar la opción FIT.

4.

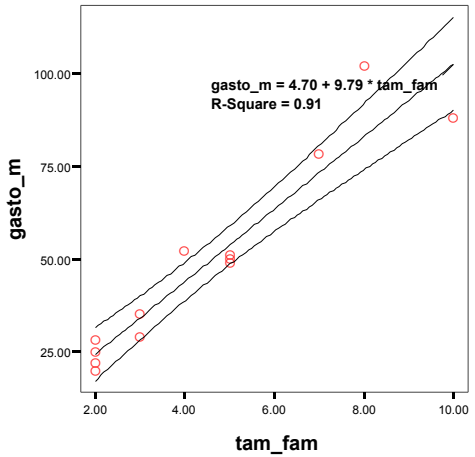




4. En la caja de diálogo, las opciones seleccionadas son: **Method** = **Regression**; se activa la opción **Include constant in equation**; en **prediction lines** se activa **Mean** y en **Fitness for** se activa **total**.

5. Se acepta y listo, tenemos una gráfica de dispersión con su línea de tendencia.

Interactive Graph



Linear Regression with
95.00% Mean Prediction Interval

Análisis del gráfico

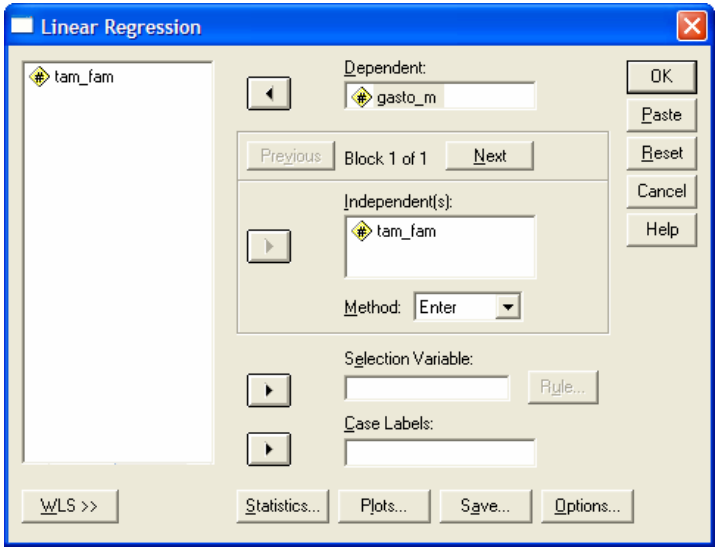
Esta gráfica muestra que si hay una tendencia lineal en los datos, la ecuación que describe esa recta y el modelo explica un 91% de la variación en los valores de Y, por efecto de los cambios en X.

Para hacer inferencias sobre los parámetros del modelo y probar los supuestos del análisis de regresión, se realiza la siguiente secuencia.

1. Seleccionar del menú:

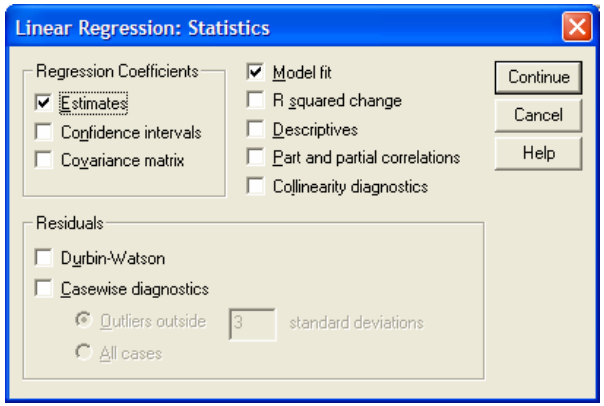
ANALYZE -> REGRESSION -> LINEAR

2. En la caja de diálogo, colocar en el lugar correspondiente la variable dependiente y la independiente, para empezar a explorar las opciones que presentan los botones ubicados en la parte inferior.



De estas opciones, las más útiles para regresión simple son STATISTICS y PLOT.

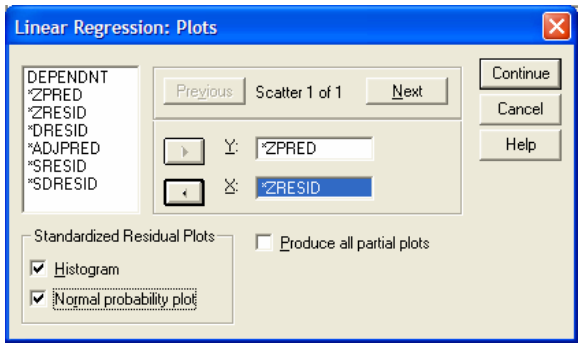
3. Abrir el diálogo STATISTICS



En este caso seleccionar ESTIMATE y MODEL FIT, para regresión múltiple utilizaremos más opciones.

4.Abrir el diálogo PLOT

Aquí se proporcionan los medios para analizar los supuestos de la prueba. Para esto seleccionar **ZPRED** (Predichos estandarizados) y **ZRESID** (Residuales estandarizados) y colocarlos en los ejes X y Y respectivamente.



Activar en Standardized Residual Plots, las opciones **HISTOGRAM** y **NORMAL PROBABILITY PLOT.**

5. Aceptar para realizar los cálculos

Resultados

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	TAM_FAM[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: GASTO_M

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.953[a]	.908	.900	8.36020

a. Predictors: (Constant), TAM_FAM

b. Dependent Variable: GASTO_M

La r^2 ajustada indica un 90% de variación explicada de la variable Y por efecto de los cambios en la variable X.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7594.254	1	7594.254	108.655	.000[a]
	Residual	768.823	11	69.893		
	Total	8363.077	12			

a. Predictors: (Constant), TAM_FAM

b. Dependent Variable: GASTO_M

ANOVA para Ho: todos los coeficientes del modelo tienen valor cero vs la Ha: al menos uno de los coeficientes del modelo es diferente de cero.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.705	4.789		.982	.347
	TAM_FAM	9.790	.939	.953	10.424	.000

a. Dependent Variable: GASTO_M

Modelo Gastos médicos mensuales = 4.705 + 9.79(Tamaño de la familia)

Residuals Statistics[a]

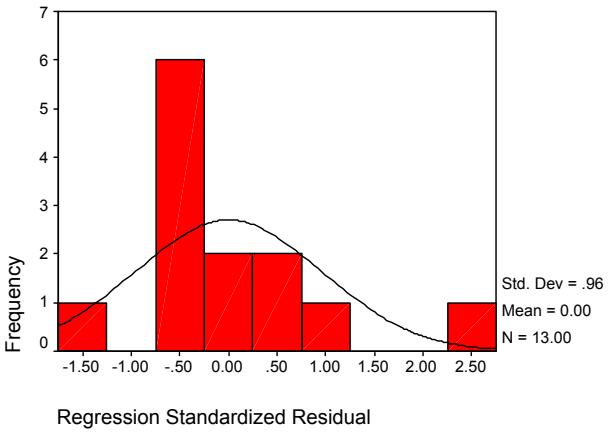
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	24.2854	102.6078	48.3846	25.15660	13
Residual	-14.6078	18.9728	.0000	8.00429	13
Std. Predicted Value	-.958	2.155	.000	1.000	13
Std. Residual	-1.747	2.269	.000	.957	13

a. Dependent Variable: GASTO_M

Charts

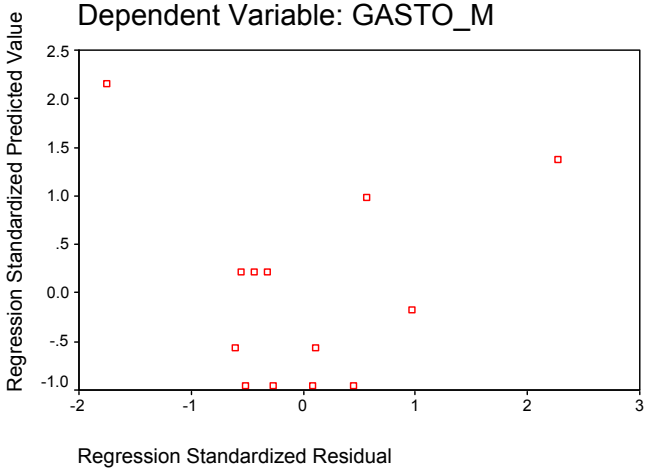
Histogram

Dependent Variable: GASTO_M



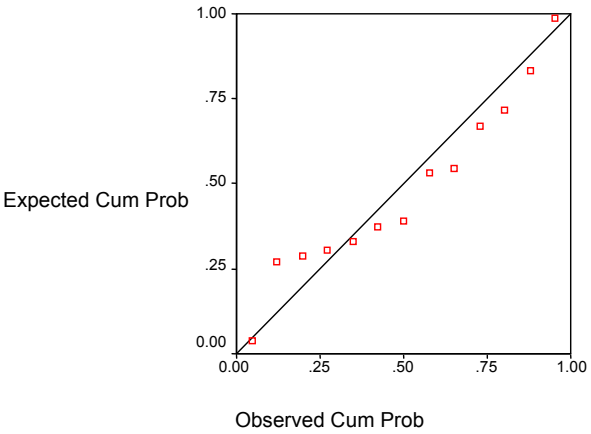
Scatterplot

Dependent Variable: GASTO_M



Normal P-P Plot of Regression Standardized Residual

Dependent Variable: GASTO_M



Análisis

El análisis de varianza muestra que al menos uno de los coeficientes del modelo es diferente de cero, en otras palabras, el modelo es significativo, entonces si hay modelo.

Después se tiene que la pendiente es diferente de cero, pero se tiene evidencia estadística de que la ordenada al origen se puede considerar cero. Entonces, se puede ajustar un modelo con ordenada igual a cero.

Las pruebas de normalidad indican que si bien los datos no son completamente normales, tampoco tienen mucha desviación de la normalidad, por lo que las conclusiones son confiables, desde el punto de vista estadístico.

Ejemplo 2. El artículo "Determination of Biological Maturity and Effects of Harvesting and Drying Conditions of Milling Quality of Paddy" (J. Agricultural Eng. Research, 1975, pp. 353-361) reporta los siguientes datos sobre la fecha X de cosecha (número de días después de la floración) y producción Y (Kg/Ha) de arroz producido en la india (**REGRESIÓN NO-LINEAL).**

(Jay L. Devore, Probabilidad y estadística para ingeniería y ciencias, 5ª. Edición, Internacional Thomson Editores, 2001, pág. 555)

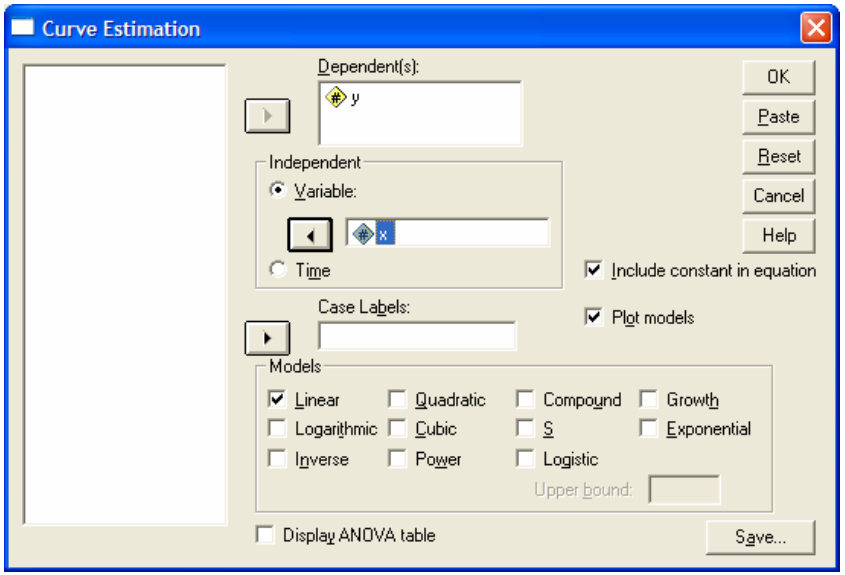
X	16	18	20	22	24	26	28	30
Y	2508	2518	3304	3423	3057	3190	3500	3883

32	34	36	38	40	42	44	46
3823	3646	3708	3333	3517	3241	3103	2776

Secuencia de análisis

1. Crear el archivo de datos con dos columnas, una para la variable independiente (X) y otra para la variable dependiente (Y).
2. Seguir la secuencia:

ANALYZE -> REGRESSION -> CURVE ESTIMATION, ya que se sospecha un comportamiento no lineal.



3. En el diálogo que se despliega, colocar las variables dependiente e independiente en el lugar que les corresponde.
   Seleccionar **LINEAR MODEL**, así como **PLOT MODELS, INCLUDE CONSTANT IN EQUATION** y **DISPLAY ANOVA TABLE**.

Resultados

Curve Fit

MODEL: MOD_1.
Dependent variable.. Y Method.. LINEAR

Listwise Deletion of Missing Data

Multiple R .27912
R Square .07791
Adjusted R Square .01205
Standard Error 415.81602

Es importante notar el valor de 1.205% para el coeficiente de determinación ajustado, muy bajo.

Analysis of Variance:

	DF	Sum of Squares	Mean Square
Regression	1	204526.2	204526.24
Residuals	14	2420641.5	172902.97

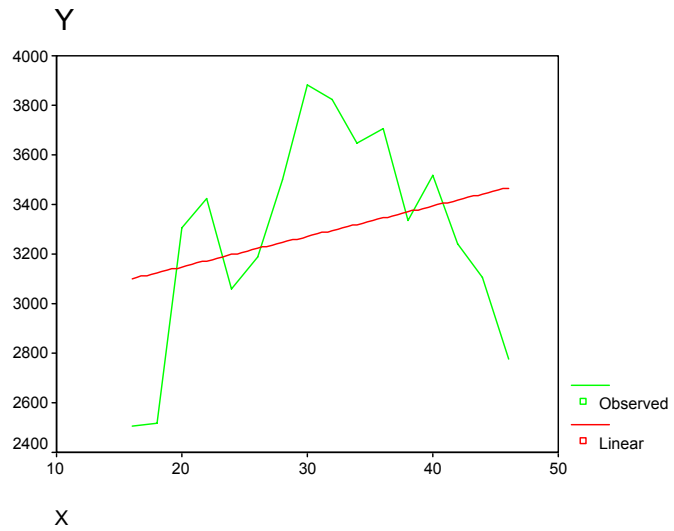
F = 1.18290 Signif F = .2951

El ANOVA muestra que todos los coeficientes son cero, entonces no hay modelo.

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	T	Sig T
X	12.263235	11.275395	.279123	1.088	.2951
(Constant)	2902.964706	364.667961		7.961	.0000

Sólo la ordenada al origen es significativa.



En este gráfico se aprecia que el modelo lineal no es la mejor opción de ajuste, ya que se presenta una curvatura, entonces se propone un modelo cuadrático.

Secuencia para un segundo modelo

Considerando los resultados del modelo anterior y siguiendo la misma secuencia se pide un modelo cuadrático

Resultados

Curve Fit
MODEL: MOD_2.

Dependent variable.. Y Method.. QUADRATI
Listwise Deletion of Missing Data

Multiple R .89115
R Square .79415
Adjusted R Square .76248
Standard Error 203.88314

Es importante el incremento de R^2 a un valor del 76.24%

Analysis of Variance:

	DF	Sum of Squares	Mean Square
Regression	2	2084779.4	1042389.7
Residuals	13	540388.4	41568.3

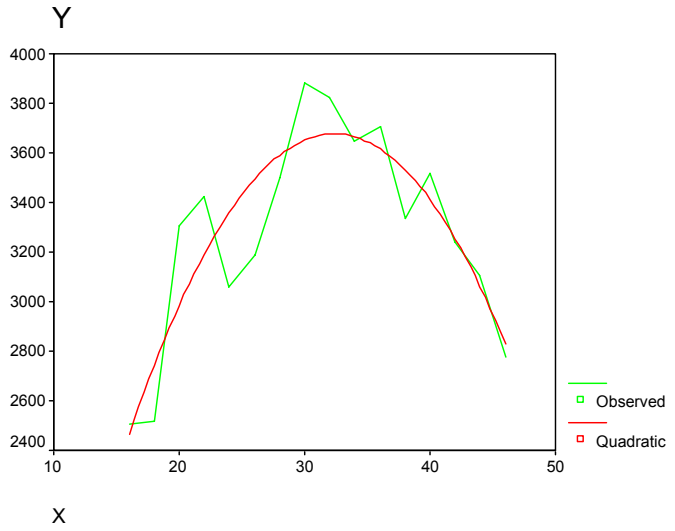
F = 25.07653 Signif F = .0000

El ANOVA muestra que al menos uno de los coeficientes del modelo es diferente de cero, entonces si hay modelo.

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	T	Sig T
X	293.482948	42.177637	6.679959	6.958	.0000
X**2	-4.535802	.674415	-6.456542	-6.726	.0000
(Constant)	-1070.397689	617.252680		-1.734	.1065

El modelo es $Y = -1070.3976 + 293.4829X - 4.5358X^2$, aunque existe evidencia de que la ordenada puede pasar por el origen, punto (0,0).



El siguiente paso es probar el modelo sin la ordenada al origen, así como verificar los supuestos del modelo estadístico.

EJEMPLO 3. (REGRESIÓN LINEAL MÚLTIPLE)

Y	X1	X2	X3
506	10	6	55
811	18	10	32
816	20	11	34
752	16	9	48
610	15	5	58
903	21	12	29
685	11	7	52
830	18	10	36
650	14	8	60
793	15	6	49
961	19	8	24
692	18	10	63
752	12	7	45
488	10	7	61
848	17	8	38
611	15	9	59
709	14	10	41
919	22	10	26
827	20	9	39
526	9	6	65

¿Hay relación lineal entre x1-x3 con y?

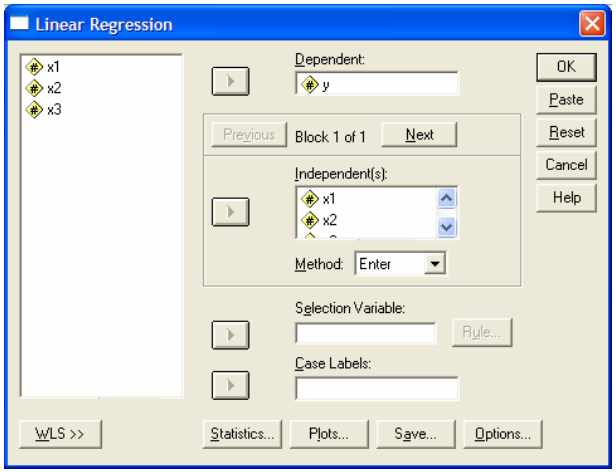
La secuencia en SPSS es:

1. Ingresar los datos en 4 columnas, una para cada variable: Y, X1, X2 y X3.

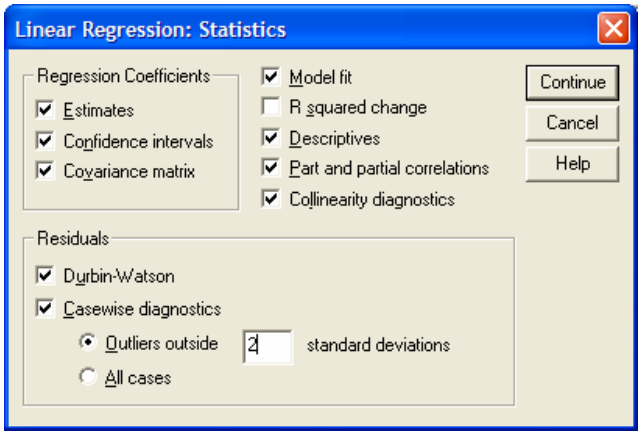
2. Seleccionar del menú:

ANALYZE -> REGRESSION -> LINEAR

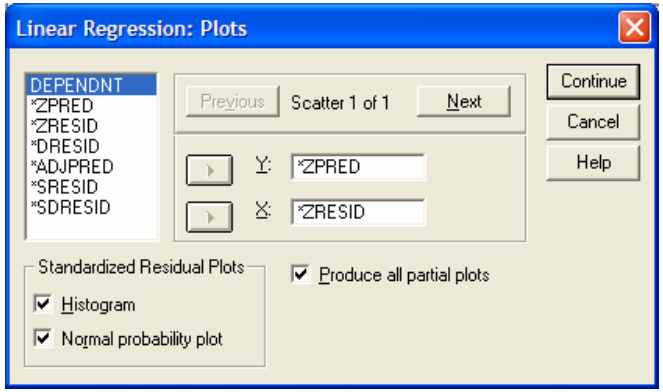
3. En la caja de diálogo, colocar en el lugar correspondiente la variable dependiente y las 3 independientes, para empezar a explorar las opciones que presentan los botones ubicados en la parte inferior.



4. En el diálogo STATISTICS, seleccionar las siguientes opciones:



5. Del diálogo PLOT seleccionar las opciones que se presentan a continuación



6. Aceptar (con un clic sobre el botón **Continue**) para realizar los cálculos

Resultados

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
Y	734.4500	137.66262	20
X1	15.7000	3.88113	20
X2	8.4000	1.90291	20
X3	45.7000	13.10725	20

En primer lugar se muestran las estadísticas descriptivas de cada variable (media, desviación estándar y número de datos), tanto dependiente como independientes.

Después se presenta una matriz de correlaciones, primero se muestran las correlaciones entre cada par de variables, después la significancia de cada correlación y por último el número de datos.

Correlations

		Y	X1	X2	X3
Pearson Correlation	Y	1.000	.847	.581	-.898
	X1	.847	1.000	.737	-.745
	X2	.581	.737	1.000	-.582
	X3	-.898	-.745	-.582	1.000
Sig. (1-tailed)	Y	.	.000	.004	.000
	X1	.000	.	.000	.000
	X2	.004	.000	.	.004
	X3	.000	.000	.004	.
N	Y	20	20	20	20
	X1	20	20	20	20
	X2	20	20	20	20
	X3	20	20	20	20

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	X3, X2, X1[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Y

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.941[a]	.886	.865	50.58898	1.592

a. Predictors: (Constant), X3, X2, X1

b. Dependent Variable: Y

El valor de R2 ajustado muestra una variación explicada del 86.5%, de manera que el modelo tiene un buen nivel explicativo.

El ANOVA muestra que al menos un coeficiente del modelo es diferente de cero, entonces se tiene evidencia estadística de que si hay modelo.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	319121.0	3	106373.678	41.564	.000[a]
	Residual	40947.917	16	2559.245		
	Total	360068.9	19			

a. Predictors: (Constant), X3, X2, X1

b. Dependent Variable: Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	837.202	127.237		6.580	.000		
	X1	17.476	5.406	.493	3.233	.005	.306	3.268
	X2	-9.961	9.046	-.138	-1.101	.287	.455	2.200
	X3	-6.421	1.330	-.611	-4.827	.000	.443	2.257

a. Dependent Variable: Y

El modelo que se obtiene es:

$$y = 837.202 + 17.476X1 - 9.961X2 - 6.421X3$$

A continuación se muestra la correlación entre los coeficientes del modelo, correlación diferente a la que se presenta entre las variables originales.

Coefficient Correlations[a]

Model			X3	X2	X1
1	Correlations	X3	1.000	.073	.575
		X2	.073	1.000	-.560
		X1	.575	-.560	1.000
	Covariances	X3	1.770	.875	4.135
		X2	.875	81.826	-27.362
		X1	4.135	-27.362	29.226

a. Dependent Variable: Y

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	X1	X2	X3
1	1	3.848	1.000	.00	.00	.00	.00
	2	.133	5.372	.00	.03	.02	.15
	3	1.322E-02	17.063	.04	.42	.97	.00
	4	5.521E-03	26.400	.96	.55	.01	.85

a. Dependent Variable: Y

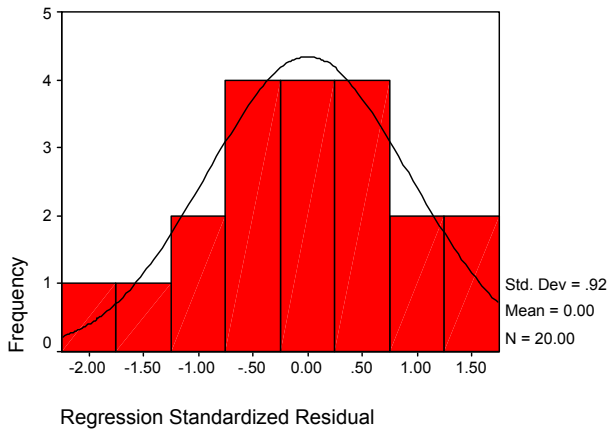
Se prueba la colinealidad del modelo y se empieza a revisar el cumplimiento de supuestos.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	617.3365	955.1105	734.4500	129.59878	20
Residual	-93.0254	68.0666	.0000	46.42363	20
Std. Predicted Val	-1.675	1.703	.000	1.000	20
Std. Residual	-1.839	1.345	.000	.918	20

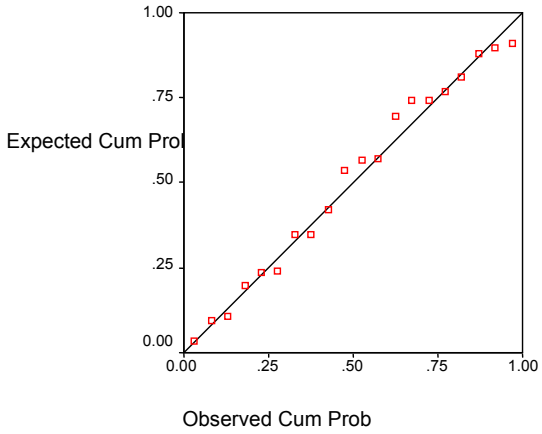
a. Dependent Variable: Y

Charts



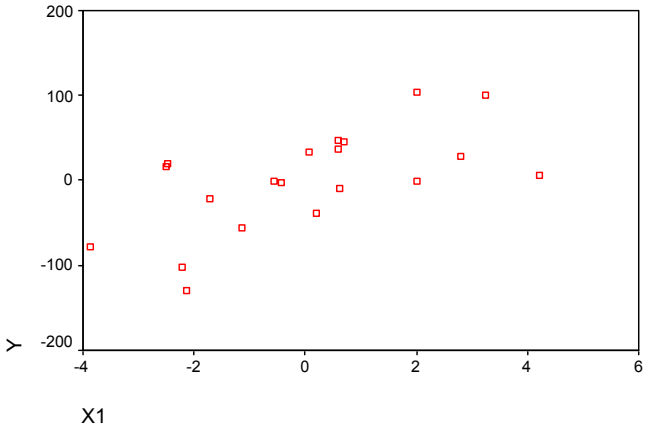
Normal P-P Plot of Regression Standardized Residual

Dependent Variable: Y



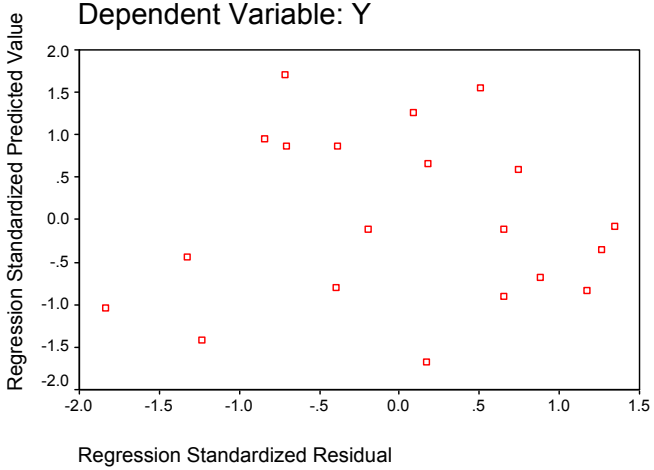
Partial Regression Plot

Dependent Variable: Y



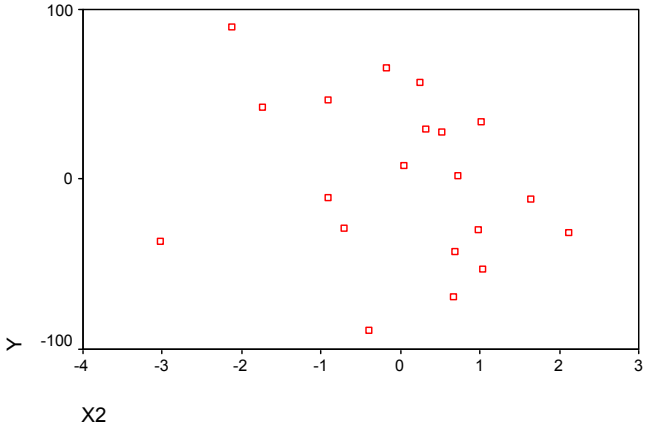
Scatterplot

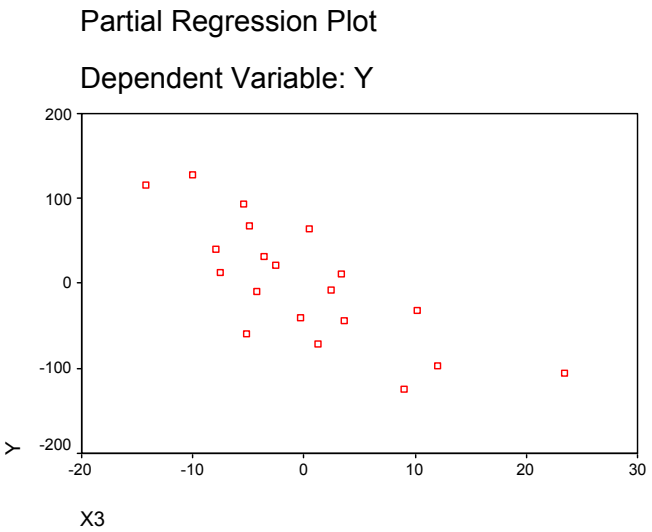
Dependent Variable: Y



Partial Regression Plot

Dependent Variable: Y





Análisis

El modelo a ajustar es del tipo

$$Y = b_o + b_1X_1 + b_2X_2 + b_3X_3$$

Del ANOVA, se rechaza Ho (Ho: $\beta_i = 0$). En otras palabras, al menos una pendiente es significativa (diferente de cero), entonces si hay modelo.

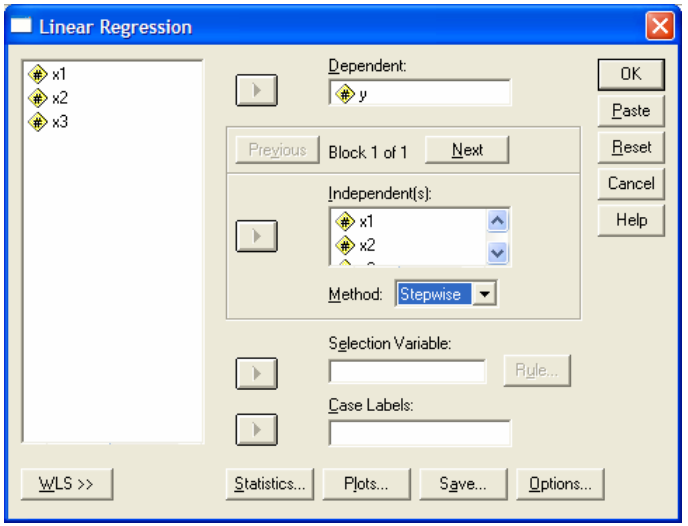
De la tabla de coeficientes se tiene el siguiente modelo:

$$y = 837.202 + 17.476X1 - 9.961X2 - 6.421X3$$

Cuyo porcentaje de variación explicada es 86.5%. En esta misma tabla se tiene que el coeficiente de X_2 es no significativo (se puede considerar cero), de tal manera que se puede obtener un mejor modelo removiendo esta variable del modelo.

Con respecto a la multicolinealidad, este problema se presenta cuando entre las variables independientes existen relaciones lineales, es decir cuando las variables independientes dependen unas de otras (unas variables son combinaciones lineales de otras). Para su detección se utiliza, en primera instancia, la matriz de correlación de coeficientes, donde se considera que hay colinealidad si el valor absoluto de la correlación es mayor a 0.75. En este caso se tiene que el valor de correlación más alto es de 0.575, **no se tiene problemas de colinealidad**. Esta conclusión se refuerza con la tabla de diagnóstico de colinealidad, donde el criterio es cuidar los eigenvalores mayores a 1, junto con la proporción de varianza. En este caso sólo el componente 1 (dimensión) tiene un eigenvalor mayor a 1, pero su proporción de varianza es cero.

Segunda opción, realizar el análisis de regresión mediante un **método de selección de variables.** Para esto seleccionar el método, de la caja de diálogo de regresión lineal **(STEPWISE)**.



También se puede seleccionar FORWARD, BACKWARD, ENTER o REMOVE, SPSS permite ingresar o remover todas las variables independientes en un solo paso, mediante estos dos últimos métodos.

Resultados

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
Y	734.4500	137.66262	20
X1	15.7000	3.88113	20
X2	8.4000	1.90291	20
X3	45.7000	13.10725	20

Estadística descriptiva de las variables

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	X3	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	X1	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: Y

Variables que "entran" o "salen" del modelo, en este caso se tiene un modelo con X_1 y X_3.

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.898[a]	.807	.796	62.16334	
2	.937[b]	.878	.863	50.90439	1.520

a. Predictors: (Constant), X3

b. Predictors: (Constant), X3, X1

c. Dependent Variable: Y

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	290511.9	1	290511.890	75.179	.000[a]
	Residual	69557.060	18	3864.281		
	Total	360068.9	19			
2	Regression	316017.6	2	158008.792	60.978	.000[b]
	Residual	44051.365	17	2591.257		
	Total	360068.9	19			

a. Predictors: (Constant), X3

b. Predictors: (Constant), X3, X1

c. Dependent Variable: Y

Poner atención en el cambio de los valores de R^2 ajustada y en el C.M de los residuales, del modelo 1 al 2.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	1165.582	51.630		22.576	.000		
	X3	-9.434	1.088	-.898	-8.671	.000	1.000	1.000
2	(Constant)	800.956	123.672		6.476	.000		
	X3	-6.315	1.335	-.601	-4.730	.000	.445	2.245
	X1	14.145	4.509	.399	3.137	.006	.445	2.245

a. Dependent Variable: Y

En la tabla anterior se tienen los valores de los coeficientes de cada uno de los modelos, mientras que en la siguiente tabla se aprecia la secuencia de variables excluidas del modelo.

Excluded Variables[c]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	X1	.399[a]	3.137	.006	.606	.445	2.245	.445
	X2	.088[a]	.684	.503	.164	.662	1.511	.662
2	X2	-.138[b]	-1.101	.287	-.265	.455	2.200	.306

a. Predictors in the Model: (Constant), X3

b. Predictors in the Model: (Constant), X3, X1

c. Dependent Variable: Y

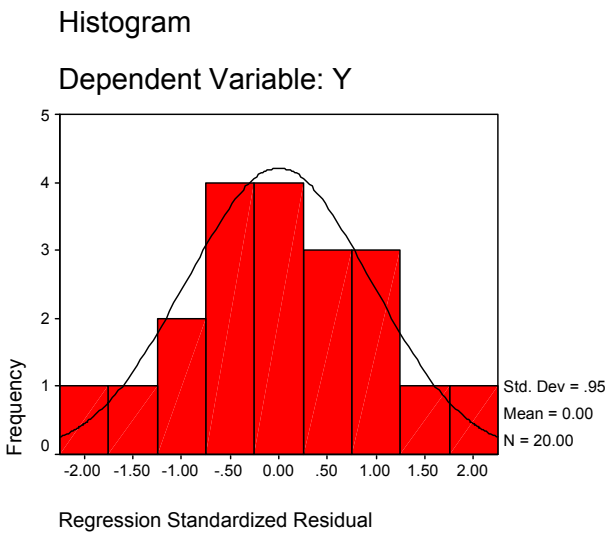
Coefficient Correlations[a]

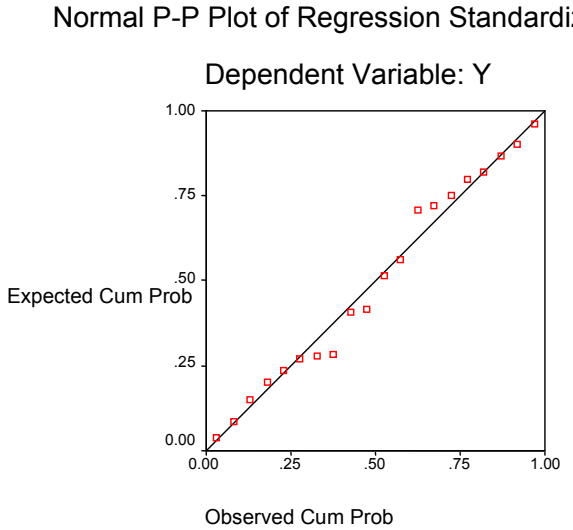
Model			X3	X1
1	Correlations	X3	1.000	
	Covariances	X3	1.184	
2	Correlations	X3	1.000	.745
		X1	.745	1.000
	Covariances	X3	1.782	4.483
		X1	4.483	20.328

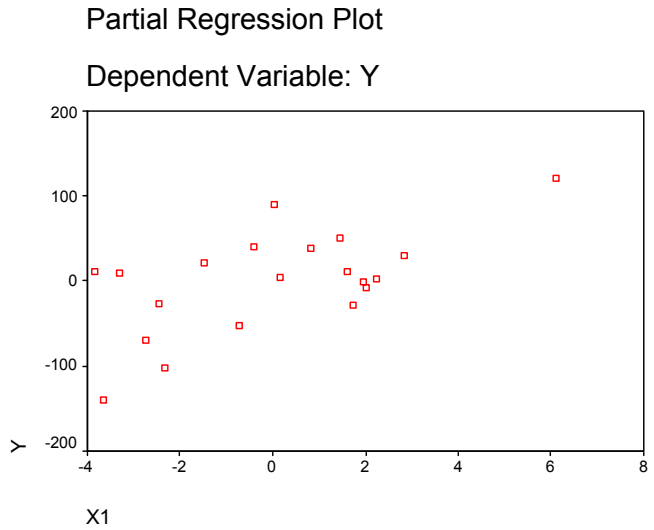
a. Dependent Variable: Y

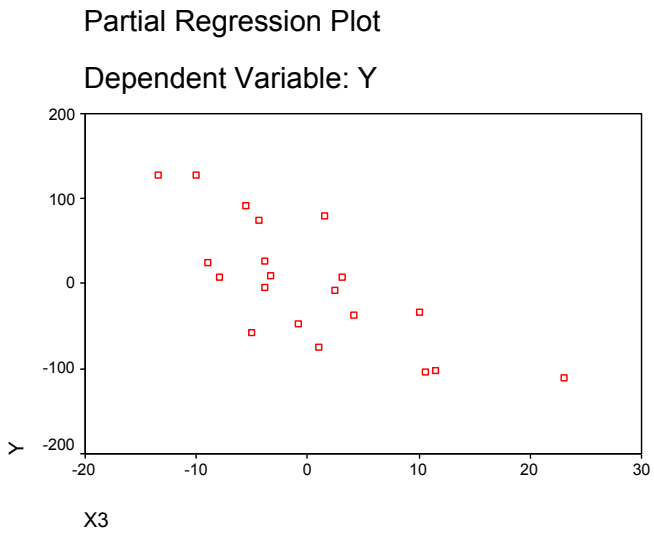
Se procede a detectar problemas de colinealidad y a probar los supuestos del modelo estadístico.

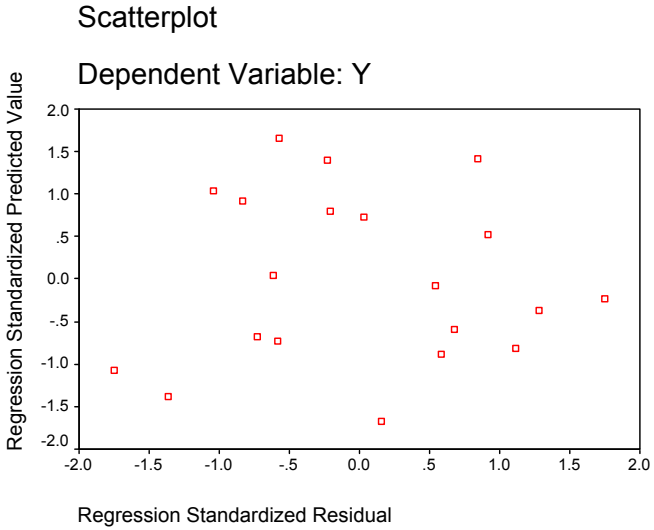
Charts











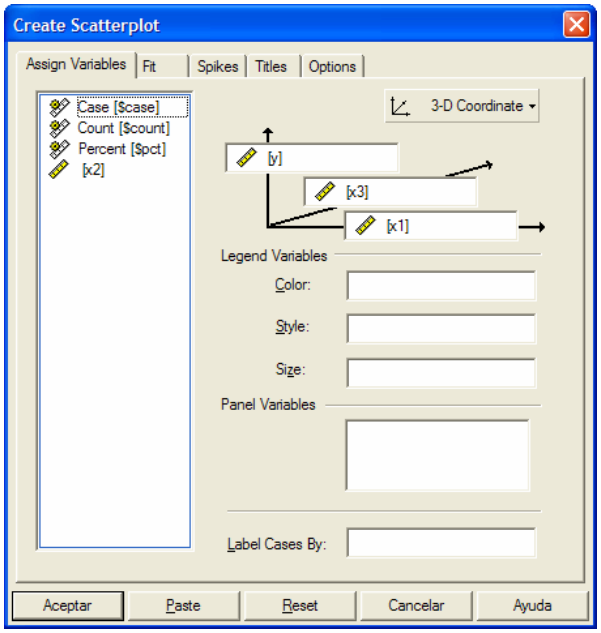
Análisis

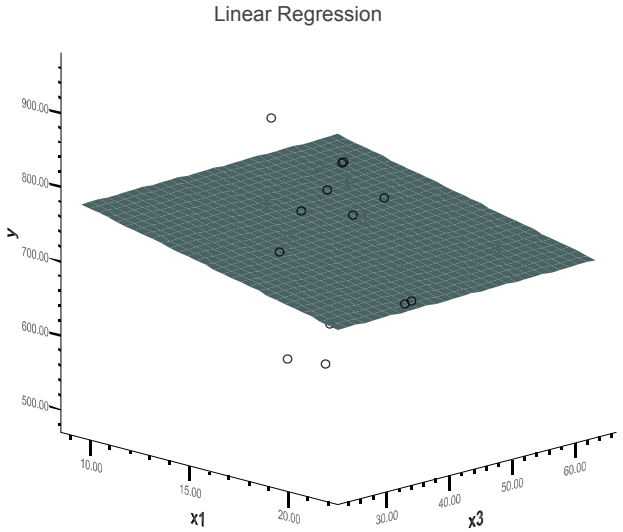
En los resultados se puede notar como cambian los valores del cuadrado medio del error y de la R^2's en este último análisis. Pero lo más notorio es como cambia el comportamiento de los residuales al verificar los supuestos de normalidad.

Hay que definir cómo queda el nuevo modelo, su porcentaje de variación explicada y la significancia de los coeficientes.

También se puede tener una gráfica de dispersión en 3 dimensiones, con la secuencia GRAPH -> INTERACTIVE -> SCATTERPLOT. Seleccionando las variables y colocándolas en el lugar adecuado.

Abrir el diálogo FIT y asegurarse que NO se seleccione alguna opción en PREDICTION LINES.





Con un clic derecho sobre la gráfica, seleccionar **SPSS Interactive Graphic Object -> EDIT** con lo cual aparece un control que permite "jugar" un poco con la gráfica para observar el modelo ajustado desde diferentes ángulos.

EJERCICIO

Y	X1	X2
100	7	28
104	11	27
106	13	29
109	15	31
115	16	26
118	18	24
123	20	20
131	23	18
136	25	22
139	28	20
150	33	19
151	34	17
153	39	14
158	41	12
159	42	14
164	44	13

¿Identificar la ecuación de regresión?

¿Hay evidencia suficiente para establecer una relación lineal positiva entre x1 con y?

¿Hay evidencia para establecer una relación lineal negativa entre x2 con y?

¿Hay suficiente evidencia para establecer que el modelo de regresión es útil?

NOTA FINAL: Un ejercicio interesante consiste en realizar un análisis de varianza y a partir de los resultados hacer un análisis de regresión. De manera que no sólo se detecten las variables significativas sino también se obtenga un modelo que describa el comportamiento de los datos.

Capítulo 6

Pruebas de Independencia y Métodos no Paramétricos

MÉTODOS NO PARAMÉTRICOS

Prueba de Kolmogorov-Smirnov

Se utiliza para contrastar las hipótesis:

Ho: La población de la cual proviene la muestra tiene una distribución normal.
Ha: La población no tiene una distribución normal.

En términos generales se utiliza para comparar un conjunto de datos continuos contra una distribución teórica, la cual puede ser NORMAL, UNIFORME, o EXPONENCIAL. Esta prueba calcula las diferencias entre los valores observados y los teóricos, analizando hasta donde las observaciones pueden razonablemente provenir de una distribución teórica especificada.

Ejemplo

Se cree que los siguientes datos provienen de una distribución normal de probabilidades. Realizar una prueba de bondad de ajuste para probar esta hipótesis.

17	23	22	24	19	23	18	22	20	13	11	21	18	20	21
21	18	15	24	23	23	43	29	27	26	30	28	33	23	29

Secuencia de análisis

1. Insertar los datos en una columna

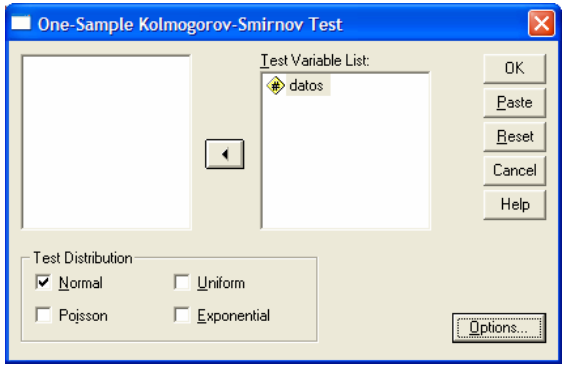
2. Seguir la secuencia

ANALYZE -> NONPARAMETRICS TESTS -> 1-SAMPLE K-S (One-Sample Kolmogorov-Smirnov Test)

3. Seleccionar la distribución teórica a comparar.

La más utilizada es la NORMAL.

Abrir el diálogo OPTIONS y seleccionar las opciones de su preferencia.



Resultados

NPar Tests

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
VAR00005	30	22.80	6.266	11	43	18.75	22.50	26.25

One-Sample Kolmogorov-Smirnov Test

		VAR00005
N		30
Normal Parameters[a,b]	Mean	22.80
	Std. Deviation	6.266
Most Extreme Differences	Absolute	.157
	Positive	.157
	Negative	-.089
Kolmogorov-Smirnov Z		.862
Asymp. Sig. (2-tailed)		.447

a. Test distribution is Normal.

b. Calculated from data.

Análisis

La significancia de 0.447, no aporta evidencia para rechazar Ho. Por lo que se tiene evidencia estadística de la normalidad de los datos.

Pruebas U-Mann-Whitney (para dos muestras independientes) y Wilcoxon (dos muestras relacionadas)

U-Mann-Whitney. Esta prueba analiza dos muestras independientes. A diferencia de la prueba de t, para dos muestras independientes, aquí lo único que se pide es que la escala de medición sea al menos ordinal, tomando como referencia de trabajo las hipótesis.

Ho: Las dos poblaciones son idénticas
Ha: Las dos poblaciones no son idénticas

¿Idénticas en qué? Pues en la variable que se esté midiendo.

Ejemplo

Los administradores de una escuela Preparatoria, recurrieron a sus registros y seleccionaron una muestra aleatoria de cuatro alumnos procedentes de una secundaria A y cinco alumnos de una secundaria B. Reportando los lugares de los alumnos dentro de su generación en la Preparatoria.

Secundaria A		Secundaria B	
Alumno	Lugar	Alumno	Lugar
A1	8	B1	70
A2	52	B2	202
A3	112	B3	144
A4	21	B4	175
		B5	146

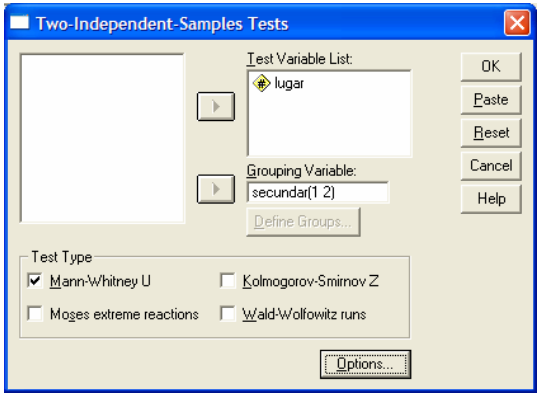
En primer lugar se ordenan los valores en una muestra combinada y se les asigna un rango del mayor al menor y se trabaja sobre ellos para contrastar las hipótesis.

Ho: Las dos poblaciones son idénticas en términos de potencial académico
Ha: Las dos poblaciones no son idénticas en términos de potencial académico

Secuencia de Análisis en SPSS

1. Ingresar los datos en dos columnas, una que identifique los dos grupos a comparar y una para la variable de respuesta.

2. Seleccionar del menú la secuencia
ANALYZE -> NONPARAMETRICS TESTS -> 2-INDEPENDENT SAMPLES



3. En la caja de diálogo, ubicar las variables en las casillas correspondientes, definiendo la variable de agrupamiento.

Resultados

NPar Tests
Mann-Whitney Test

Ranks

	SECUNDAR	N	Mean Rank	Sum of Ranks
LUGAR	1	4	2.75	11.00
	2	5	6.80	34.00
	Total	9		

Test Statistics[b]

	LUGAR
Mann-Whitney U	1.000
Wilcoxon W	11.000
Z	-2.205
Asymp. Sig. (2-tailed)	.027
Exact Sig. [2*(1-tailed Sig.)]	.032[a]

a. Not corrected for ties.

b. Grouping Variable: SECUNDAR

Análisis

De acuerdo a los valores de significancia (menores a 0.05) se tiene evidencia para rechazar Ho, y además se puede ver en los rangos promedios que los alumnos de la secundaria 2 tienen mejores lugares en la preparatoria, por lo que se puede concluir que tienen mayor potencial académico que los de la secundaria 1.

Prueba de Rangos con signos de Wilcoxon

Esta es la prueba análoga a la comparación de muestras pareadas. Se basa en obtener las diferencias absolutas entre las dos muestras a comparar y ordenarlas de menor a mayor, después asignarles un rango, cuyo signo corresponde al signo de la diferencia original y a partir de estos datos contrastar las hipótesis.

Ho: Las dos poblaciones son idénticas
Ha: Las dos poblaciones no son idénticas

Ejemplo

Una fábrica trata de determinar si dos métodos de producción tienen distintos tiempos de terminación de lote. Para esto se selecciona una muestra de 11 trabajadores y cada uno terminó un lote de producción usando los dos métodos. Seleccionando el método inicial para cada trabajador de manera aleatoria.

Los datos son tiempo de terminación de lote (minutos).

Trabajador	1	2	3	4	5	6	7	8	9	10	11
Método1	10.2	9.6	9.2	10.6	9.9	10.2	10.6	10.0	11.2	10.7	10.6
Método 2	9.5	9.8	8.8	10.1	10.3	9.3	10.5	10.0	10.6	10.2	9.8

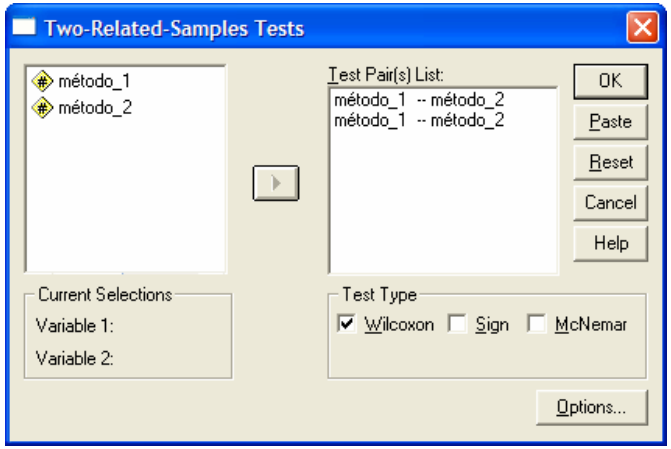
Secuencia de análisis

1. Al igual que en la prueba paramétrica, insertar los datos en dos columnas, una para la condición inicial del estudio y otra para la final.

2. Del menú seguir la secuencia

ANALYZE -> NONPARAMETRICS TESTS -> 2-RELATED-SAMPLE

3. En la caja de diálogo, ubicar las variables en las casillas correspondientes, definiendo las variable a comparar.



4. Seleccionar el tipo de prueba, en este caso Wilcoxon.

Resultados

NPar Tests
Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
MÉTODO_2 - MÉTODO_1	Negative Ranks	8[a]	6.13	49.00
	Positive Ranks	2[b]	3.00	6.00
	Ties	1[c]		
	Total	11		

a. MÉTODO_2 < MÉTODO_1
b. MÉTODO_2 > MÉTODO_1
c. MÉTODO_1 = MÉTODO_2

Test Statistics[b]

	MÉTODO_2 - MÉTODO_1
Z	-2.193[a]
Asymp. Sig. (2-tailed)	.028

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

Análisis

Se tiene evidencia para rechazar Ho, por la significancia menor al 0.05 (en este caso 0.028), por lo tanto los dos métodos no son idénticos en cuanto a su tiempo de terminación de lote.

Prueba de Kruskal-Wallis

La prueba de Mann-Whitney se aplica cuando se tienen dos poblaciones a comparar y Kruskal y Wallis cuando se tienen tres o más poblaciones. El par de hipótesis a trabajar es.

Ho: Todas las poblaciones son idénticas
Ha: NO todas las poblaciones son idénticas

La prueba de Kruskal-Wallis se puede utilizar con datos ordinales y también con datos de intervalos o de relación. No requiere supuestos de normalidad ni de homogeneidad de varianza.

El estadístico de contraste es

$$W = \left(\frac{12}{n_T (n_T + 1)} \sum_{i=1}^{k} \frac{R_1^2}{n_i} \right) - 3(n_T + 1)$$

donde

k = número de poblaciones a comparar
n_i = número de repeticiones en la muestra i
$n_T = \Sigma n_i$ =número total de repeticiones en todas las muestras
R_i = suma de los rangos en la muestra i

Se ha demostrado que la W de Kruskal Wallis se aproxima a una chi-cuadrada, con k-1 grados de libertad, bajo el supuesto de que se cumpla Ho. Aproximación que se cumple mejor cuando el número de repeticiones de cada muestra es mayor o igual a 5.

Como se ve en la fórmula de W, sólo se trabaja con Rangos de manera que se deben ordenar todo los datos de menor a mayor y asignarles rangos. Sin perder la identificación del grupo o tratamiento al que pertenece cada valor.

Ejemplo

Tres productos recibieron las siguientes calificaciones por parte de un jurado de 15 consumidores.

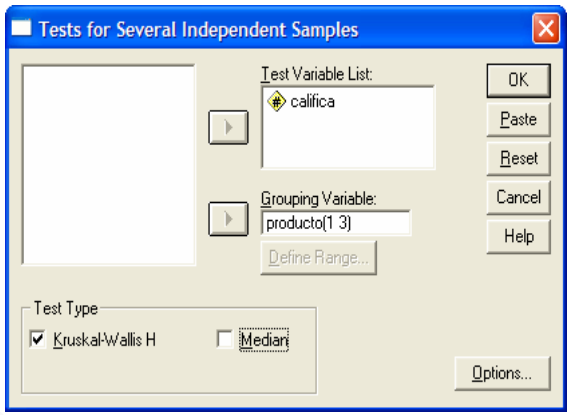
Producto A	50	62	75	48	65
Producto B	80	95	98	87	90
Producto C	60	45	30	58	57

¿Hay diferencia apreciable en las calificaciones de evaluación de los productos?

Secuencia de Análisis

1. Ingresar los datos en dos columnas, una para los identificadores de grupos o tratamientos y otra para los resultados o variable dependiente.

2. Del menú seguir la secuencia
ANALYZE -> NONPARAMETRICS TESTS -> K-INDEPENDET-SAMPLES



3. En la caja de diálogo, ubicar las variables en las casillas correspondientes, definiendo la variable de agrupamiento.

Resultados

NPar Tests
Kruskal-Wallis Test

Ranks

	PRODUCTO	N	Mean Rank
CALIFICA	Producto A	5	6.80
	Producto B	5	13.00
	Producto C	5	4.20
	Total	15	

Test Statistics[a,b]

	CALIFICA
Chi-Square	10.220
df	2
Asymp. Sig.	.006

a. Kruskal Wallis Test

b. Grouping Variable: PRODUCTO

Análisis

Se tiene evidencia para decir que no todas las evaluaciones son idénticas, ya que la significancia es 0.006. En este caso por los valores de los rangos promedios, se podría pensar que el producto B es el que tiene una calificación más alta.

Este análisis se puede complementar con diagramas de cajas o con la opción explore.

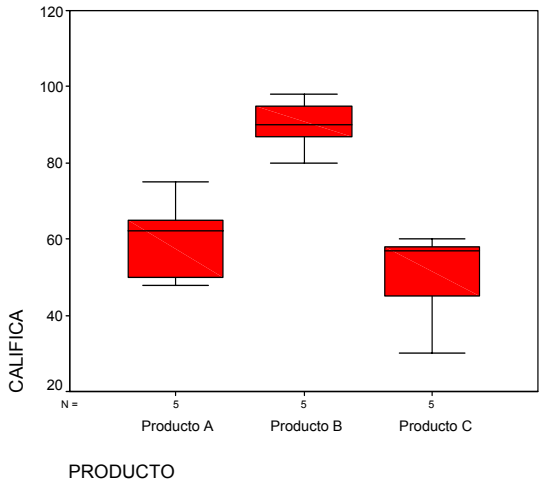
Explore

CALIFICA

Stem-and-Leaf Plots

```
CALIFICA Stem-and-Leaf Plot for
PRODUCTO= Producto A

 Frequency    Stem &  Leaf
     1.00       4 .  8
     1.00       5 .  0
     2.00       6 .  25
     1.00       7 .  5

 Stem width:        10
 Each leaf:       1 case(s)


CALIFICA Stem-and-Leaf Plot for
PRODUCTO= Producto B

 Frequency    Stem &  Leaf
     1.00       8 .  0
     1.00       8 .  7
     1.00       9 .  0
     2.00       9 .  58

 Stem width:        10
 Each leaf:       1 case(s)


CALIFICA Stem-and-Leaf Plot for
PRODUCTO= Producto C

 Frequency    Stem &  Leaf
     1.00       3 .  0
     1.00       4 .  5
     2.00       5 .  78
     1.00       6 .  0

 Stem width:    10
 Each leaf:
1 case(s)
```

Los gráficos de tallos y hojas permiten comparar la distribución de los datos en cada uno de los productos, ya que este tipo de pruebas son más confiables cuando los datos tienen una distribución semejante. Y el boxplot permite visualizar las diferencias en las calificaciones.



PRUEBA DE INDEPENDENCIA

La Chi-cuadrada se puede utilizar para comparar una varianza con un valor dado, realizar pruebas de bondad de ajuste (¿hasta dónde una muestra se comporta de acuerdo a una distribución dada?), pero mucha de su popularidad viene de la posibilidad de **probar la independencia** o relación entre dos variables, generalmente de tipo categórico, y arregladas en tablas de doble entrada con **r** filas o renglones y **c** columnas, a la cual se le conoce como tabla de contingencia.

Pasos para hacer una prueba de contingencia:

1. Plantear las hipótesis a contrastar

Ho: La variable de la columna es independiente de la variable del renglón
Ha: La variable columna NO es independiente de la variable renglón

En términos coloquiales: **Ho: NO hay relación entre las variable**.

2. Tomar una muestra aleatoria y anotar las frecuencias observadas para cada celda de la tabla de contingencias

3. Aplicar la siguiente ecuación: $E_{i,j} = (\Sigma r_i)(\Sigma c_j)/N$ (total del renglon i multiplicado por el total de la columna j y dividir este resultado entre el total o tamaño de la muestra), para calcular las frecuencias esperadas en cada celda.

4. Obtener un valor para el estadístico chi- cuadrada

$$\chi^2 = \Sigma [(O - E)^2 / E]$$

una medida de la desviación entre las frecuencias observadas y las esperadas.

5. Aplicar la regla de decisión.

Manejo numérico.

Tabla de Contingencia, valores observados

	Col_1	*Col_2*	*Col_3*	*Totas*
Reng_1	12	2	6	20
Reng_2	6	6	8	20
Total	18	8	14	

Tabla de Contingencia, valores esperados

	Col_1	*Col_2*	*Col_3*	*Total*
Reng_11	$O = 12$ $E = 9$	$O = 2$ $E = 4$	$O = 6$ $E = 7$	$\Sigma O = 20$ $\Sigma E = 20$
Reng_2	$O = 6$ $E = 9$	$O = 6$ $E = 4$	$O = 8$ $E = 7$	$\Sigma O = 20$ $\Sigma E = 20$
Total	$\Sigma O = 18$ $\Sigma E = 18$	$\Sigma O = 8$ $\Sigma E = 8$	$\Sigma O = 14$ $\Sigma E = 14$	

Para este caso: $\chi^2 = \Sigma [(O - E)^2 / E] = 30/7 = 4.3$

Con grados de libertad g.l. = (2-1)(3.1) = 2, se tiene una significancia de 0.14, lo que implica NO rechazar Ho.

Una de las primeras preguntas es que tan grande es este valor de chi-cuadrada, por lo que el valor máximo se obtiene con: $\chi^2_{max} = N(A-1)$, con A es valor más pequeño de las hileras o columnas. En este caso el valor numérico es: 40(2-1) = 40(1) = 40.

De aquí se puede obtener el coeficiente de determinación, como: χ^2/χ^2_{max} = 4.3/40 = 0.11. De donde se concluye que el grado de asociación es de únicamente un 11%.

Ejemplo 1. Se tiene un estudio para probar si la preferencia del tipo de cerveza (ligera, clara u oscura), es independiente del sexo del consumidor (hombre o mujer), encontrando los siguientes resultados.

	Ligera	Clara	Oscura	Total
Hombre	20	40	20	80
Mujer	30	30	10	70
Total	50	70	30	150

Secuencia de análisis en SPSS

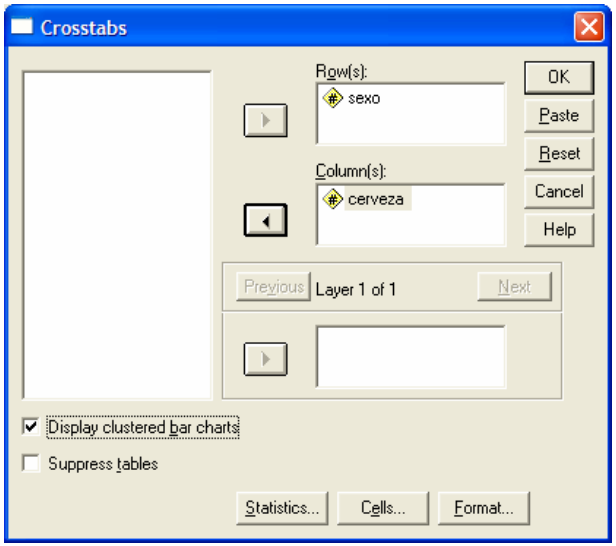
1. Ingresar los datos pero no como concentrados, sino como observaciones. De tal manera que para este ejemplo se tienen 150 datos. Una columna para sexo o género (que puede tomar 2 valores) y otra para tipo de cerveza (con tres posibles valores).

2. Seguirla secuencia

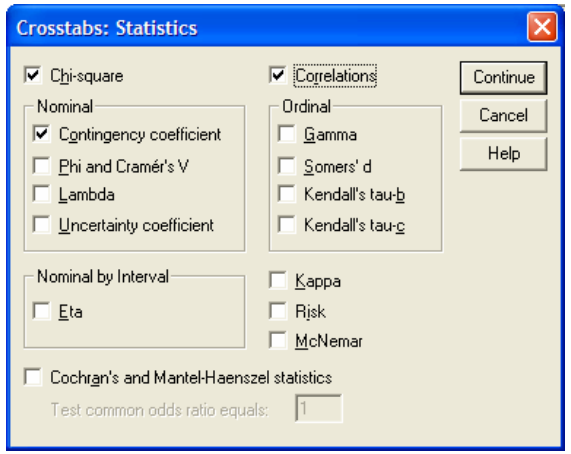
ANALYZE -> DESCRIPTIVE STATISTICS -> CROSSTAB

3. Seleccionar las opciones correspondientes

Como pedir que despliegue un gráfico de barras.



4. Abrir el diálogo de estadísticas



Resultados

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
SEXO * CERVEZA	150	100.0%	0	.0%	150	100.0%

SEXO * CERVEZA Crosstabulation

			CERVEZA			Total
			ligera	clara	oscura	
SEXO	Hombres	Count	20	40	20	80
		Expected Count	26.7	37.3	16.0	80.0
	Mujeres	Count	30	30	10	70
		Expected Count	23.3	32.7	14.0	70.0
Total		Count	50	70	30	150
		Expected Count	50.0	70.0	30.0	150.0

Tabla de contigencia generada a partir de los datos ingresados.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6.122[a]	2	.047
Likelihood Ratio	6.178	2	.046
Linear-by-Linear Association	5.872	1	.015
N of Valid Cases	150		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 14.00.

Prueba de Ji-cuadrada con sus valores de Significancia.

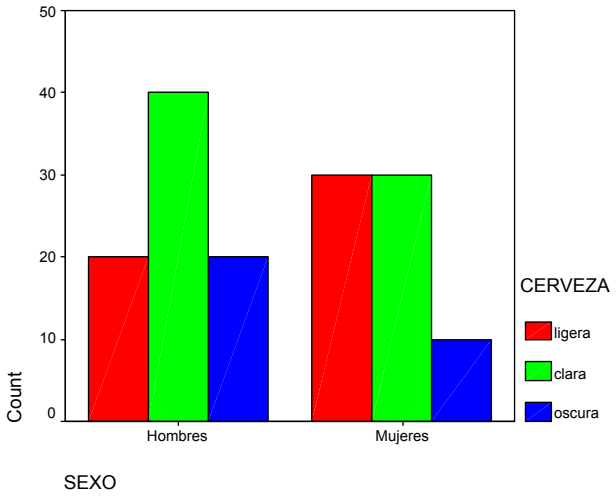
Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Contingency Coefficient	.198			.047
Ordinal by Ordinal	Kendall's tau-b	-.190	.075	-2.526	.012
	Spearman Correlation	-.200	.079	-2.489	.014[c]
Interval by Interval	Pearson's R	-.199	.079	-2.464	.015[c]
N of Valid Cases		150			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.



Análisis

Revisar las significancias a la luz de la hipótesis de trabajo. Por ejemplo, el valor de Chi-cuadrada está por abajo del 0.05 por lo que se podría rechazar la Ho para la prueba de independencia. Se pueden aprovechar también los resultados del coeficiente de correlación y entonces se tiene que el juego de hipótesis es:

Ho: $\rho = 0$ (no hay correlación)
Ha: ρ diferente de cero (si hay correlación)

Donde puede verse que si hay correlación entre ambas variables.

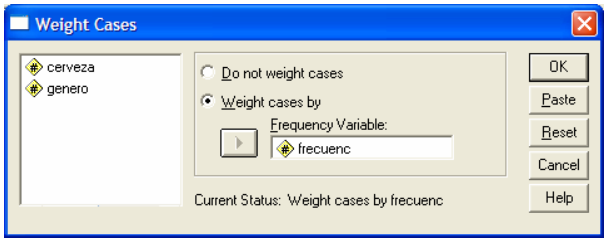
"Teclear" o ingresar los datos de esta manera es bastante engorroso, así que se tiene una forma alternativa de ingresarlos ya condensados.

Para esto se "acomodan" los datos en una tabla con el siguiente formato.

Cerveza	Género	Frecuenc
1	1	20
1	2	30
2	1	40
2	2	30
3	1	20
3	2	10

El archivo de datos consta, entonces, de 3 columnas y sólo 6 filas. Antes de realizar el análisis de la manera ya descrita se aplica la secuencia.

DATA -> WEIGHT CASES



Asegurándose de que esté seleccionada la opción **Weight cases by** y que la variable frecuenc quede en **Frequency Variable**.

Ejemplo 2. Una encuesta sobre el deporte preferido tuvo los siguientes resultados en hombres y mujeres

	Béisbol	Basquetbol	Futbol	Total
Hombre	19	15	24	58
Mujer	16	18	16	50
Total	35	33	40	108

¿Son iguales las preferencias entre hombres y mujeres?

Realizando el análisis con la secuencia y descrita, se tienen los siguientes resultados.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
SEXO * DEPORTE	108	100.0%	0	.0%	108	100.0%

SEXO * DEPORTE Crosstabulation

Count

		DEPORTE			Total
		Beisbol	Basquetbol	Futbol	
SEXO	Hombres	19	15	24	58
	Mujeres	16	18	16	50
Total		35	33	40	108

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1.546[a]	2	.462
Likelihood Ratio	1.548	2	.461
Linear-by-Linear Association	.286	1	.593
N of Valid Cases	108		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 15.28.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nomina	Contingency Coefficie	.119			.462
Ordinal by Ordinal	Kendall's tau-b	-.051	.090	-.561	.575
	Spearman Correlation	-.054	.096	-.554	.580[c]
Interval by Interval	Pearson's R	-.052	.096	-.533	.595[c]
N of Valid Cases		108			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

Análisis. No se rechaza Ho.

Bibliografía

Box George E. P., William G. Hunter y J. Stuart Hunter, 1999, ESTADÍSTICA PARA INVESTIGADORES. Introducción al Diseño de Experimentos, Análisis de Datos y Construcción de Modelos, Ed. Reverté, México.

Byrman A. y Duncan Cramer, 2001, QUANTITATIVE DATA ANALYSIS WITH SPSS RELEASE 10 FOR WINDOWS. A guide for social scientists, Ed. Routledge, USA.

Camacho R. J., 2001, ESTADÍSTICA CON SPSS Ver. 9 PARA WINDOWS, Ed. Alfaomega, grupo editor Ra-Ma, México.

Carver Robert H. and Jane Gradwohl Nash, 2000, DOING DATA ANALYSIS WITH SPSS 10.0, Dexbury Thomson Learning, Canadá.

Devore L. J., 2001, PROBABILIDAD Y ESTADÍSTICA PARA INGENIERÏA y CIENCIAS, International Thomson editores, México.

Freund John E. y Gary A. Simon, 1992, ESTADÍSTICA ELEMENTAL, Ed. Prentice Hall, México.

Lizasoain L. y Luis Joaristi, 1996, SPSS PARA WINDOWS Ver. 6.01 EN CASTELLANO, Ed. Paraninfo, México.

Montgomery Douglas C., 2002, DISEÑO Y ANÁLISIS DE EXPERIMENTOS, 2ª ed., Ed. Limusa-Wiley, México.

Neter J., Kutner H. M., Wasserman W., 1996, APPLIED LINEAR REGRESSION MODELS, Ed. Times Mirror Higher Education Group, E.U.A.

Wonnacott T. H., y R. J. Wonnacott, 1999, INTRODUCCIÓN A LA ESTADÍSTICA, 5ª. Ed. del inglés, Ed. Limusa, Grupo Noriega, México.

Contenido

Capítulo 1

Conociendo el entorno de trabajo SPSS versión 11
Se describe el entorno de trabajo de *SPSS*, enfatizando en el manejo del editor de datos, se explica como guardar archivos de datos y de resultados en disco y se empieza a mostrar el análisis con las opciones gráficas.

Capítulo 2

Describiendo los datos
Se explica como obtener las estadísticas descriptivas de un conjunto de datos, con la opción *descriptive y explore*, enfatizando en la combinación de gráficos y resultados numéricos para describir y/o explorar los datos, antes de aplicar cualquier análisis inferencial.

También se explican una serie de conceptos y términos necesarios para definir y seleccionar la técnica estadística más adecuada y para apoyar la interpretación de resultados.

Capítulo 3

Introducción a la Inferencia
Se explica brevemente en que consiste la inferencia estadística, se dan las fórmulas de cálculo de intervalos de confianza y pruebas de hipótesis para las medias o varianzas (una media o varianza contra un valor definido de antemano y la comparación de un par de medias o varianzas; mostrando la secuencia a seguir para realizar este tipo de análisis en SPSS.

Capítulo 4

Análisis de Varianza y Diseño de Experimentos
Se hace una breve revisión del concepto de Análisis de Varianza, mostrando la estrategia de cálculo numérico. A partir de ahí se muestra lo que es un diseño de una-vía o completamente al azar, un diseño de bloques al azar y los diseños factoriales de dos vías (dos factores). Enfatizando en la verificación de supuestos para garantizar la validez de las conclusiones. En los ejercicios se plantea un diseño cuadrado latino y un factorial con tres factores.

Capítulo 5

Análisis de Regresión
Se analiza en que consiste el método de mínimos cuadrados, tanto para la regresión lineal simple como para la múltiple. Se explica brevemente como se interpreta el *ANVA* para un análisis de regresión y las pruebas de hipótesis para los coeficientes de un modelo lineal.

También se revisan los métodos de selección de variables y los criterios para definir cuando un modelo es mejor que otro.

Capítulo 6

Pruebas de Independencia y Métodos no Paramétricos
Se revisan las pruebas No-Paramétricas análogas a las revisadas en el capítulo 3 y 4. Así como una prueba de bondad de ajuste y el clásico análisis de Ji-Cuadrada para probar la independencia entre las variables de una tabla de doble entrada. También se revisa la correlación no paramétrica.

Bibliografía
Se dan algunas referencias bibliográficas que pueden apoyar tanto el aspecto estadístico, como el uso del "paquete" SPSS. Material relativamente accesible por el nivel técnico de la información que presentan y por su disponibilidad en las bibliotecas

PAPIME: EN216403

FES ZARAGOZA, UNAM, 2004

ISBN: 970-32-1260-3