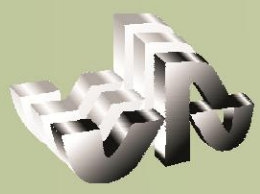
OSIRIS CASTEJÓN SANDOVAL

DISEÑO Y ANÁLISIS DE EXPERIMENTOS CON STATISTIX



Universidad Rafael Urdaneta

FONDO EDITORIAL BIBLIOTECA

Diseño y análisis de experimentos con Statistix

Osiris Castejón Sandoval

Diseño y análisis de experimentos con Statistix

FONDO EDITORIAL BIBLIOTECA
Universidad Rafael Urdaneta

Universidad Rafael Urdaneta

Autoridades Rectorales

Dr. Jesús Esparza Bracho, Rector
Ing. Maulio Rodríguez, Vicerrector Académico
Ing. Salvador Conde, Secretario

Portada: Luz Elena Hernández

Universidad Rafael Urdaneta, Fondo Editorial Biblioteca
Vereda del Lago, Maracaibo, Venezuela.

ISBN: 978-980-7131-14-8

Deposito Legal: lfi23820116202737

Dedicatoria

A mis hijos:
Marielba Alejandra
Orlando José
Carlos Alberto
Julia Gabriela
Magalys María

El tesoro más preciado que Dios me dio

Tabla de Contenido

Introducción
Capítulo I. Principios básicos del diseño experimental 1
Capítulo II. Diseño completamente aleatorizado 7
Capítulo III. Diseño en bloques completos al azar 27
Capítulo IV. Diseño de cuadrado latino y diseños afines 39
Capítulo V. Experimentos factoriales 49
Capítulo VI. Análisis de covarianza 69
Capítulo VII. Análisis de varianza no paramétricos 82
Capítulo VIII. Análisis de correlación simple, múltiple, parcial 91
Capítulo IX. Análisis de regresión no lineal 120
Capítulo X. Regresión logística 137
Capítulo XI. Regresión theil (Regresión no paramétrica) 148
Capítulo XII. Regresión con variables ficticias 152
Apéndice A 161
Apéndice B 176
Bibliografía 198

Introducción

El presente libro está diseñado para aprender a resolver problemas de investigación con diseños experimentales fundamentados en el enfoque Fisheriano.

Está estructurado en doce capítulos, los primeros seis dedicados a revisar los diseños experimentales y de tratamiento, propiamente dicho, y, los restantes a revisar los contenidos relacionados con el análisis de regresión y correlación.

El enfoque que se sigue a lo largo del libro es intuitivo, con poca o casi ninguna demostración estadístico matemática.

Se expone en forma breve el contenido teórico ligado a cada diseño, lo fundamental, y se hace hincapié en el análisis informatizado utilizando el software STATISTIX versión 8 para Windows.

En forma esquemática se siguen estos pasos:

- Enunciado del problema
- Hipótesis estadísticas
- Preparación de la matriz de base de datos
- Transcripción en el software estadístico
- Procesamiento
- Análisis de las salidas del software
- Interpretación de los resultados y
- Conclusiones

Se presentan dos alternativas de análisis el paramétrico y no paramétrico. El primero de uso predilecto en el caso de que los supuestos que fundamentas las pruebas estadísticas se cumplan y el segundo para cuando no se cumplan éstos supuestos o el investigador no desee hacer ningún tipo de suposición.

El software estadístico puede ser bajado del sitio web:

http://www.statistix.com

El autor espera que la recopilación presentada aunado a su dilatada experiencia que arriba a más de veinte años de ejercicio docente pueda llenar las expectativas que en los investigadores nóveles y otros despiertan éstos tópicos de amplia aplicación en el campo de la investigación científica.

Osiris Castejón Sandoval
Profesor Titular Jubilado de LUZ
Magíster en Informática Educativa

Capitulo I

Principios Básicos del Diseño Experimental

Hace más de seis décadas que Sir Ronald A. Fisher, sentó los cimientos que ha llegado a conocerse por el título de su libro "The design of experiment". Desde entonces la teoría del diseño de experimento ha sido desarrollada y ampliada considerablemente.

La experimentación proporciona lo que se conoce por datos experimentales, en contraste con las informaciones que se toman en la ciencia no experimental, que se conocen como datos de observación. Los datos de observación provienen de las observaciones directas sobre los elementos de una población o muestra, y no deberían ser cambiados ni modificados en el transcurso de la investigación.

A menudo, es difícil asignar causa y efecto por el estudio de los datos de observación. Si el interés es establecer relaciones causales, se debe trabajar con datos experimentales, obtenidos de observaciones de un conjunto o segmento de él, donde se han controlado o modificado ciertos factores, si es que ejercen algún efecto sobre los datos. En otras palabras, los datos experimentales son el resultado de experimentos diseñados lógicamente, que ofrecen pruebas a favor o en contra de las teorías de causa y efecto.

Merecen citarse algunos ejemplos básicos en el diseño experimental:

(a)Experimentos donde se investiga una relación de causa-efecto: En este tipo de experimentos se estudia el efecto de varias variables independientes (TRATAMIENTOS) en una respuesta (variable dependiente). Las variables independientes suelen llamarse: tratamientos o factores, que pueden ser de naturaleza cuantitativa o cualitativa. Se supone que los valores de una respuesta refleja los valores de distintos tratamientos.

(b)Experimentos controlados: La experimentación generalmente consiste en experimentos controlados, en lo que hace variar uno o más factores mientras que los demás factores relevantes permanecen constantes; sin embargo, con frecuencia no es posible controlar solamente unos factores exactamente, excepto aquellos cuyo efecto esta siendo investigados, por consiguiente, existe siempre un error experimental, o la variación respuesta, debido a la falta de control preciso. El error experimental se mide por la variación de observaciones, hechas en condiciones supuestamente idénticamente.

Cuando se realizan experimentos controlados en el mundo real, en el que innumerables factores actúan en diversos grados de influencia, los resultados serán mucho más variables. En este caso, es conveniente considerar toda la varianza que no puede ser explicada por la variación de los tratamientos experimentales, como error experimental, parte del cual podría ser explicado si el tratamiento se hubiera realizado en condiciones ideales. Debido a esto, los tratamientos en el mundo real tienden a sobreestimar el error experimental.

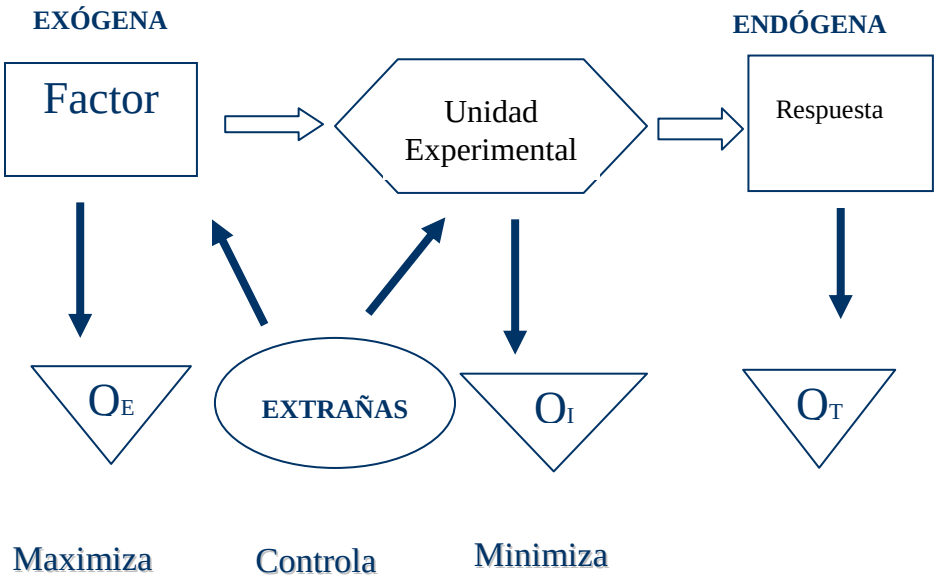
Diseño experimental

Es una secuencia de pasos tomados de antemano (planeado) para asegurar que los datos se obtendrán adecuadamente, lo que permitirá un análisis objetivo conducente a conclusiones válidas del problema Investigado.

Objetivo

Es el arreglo de unidades experimentales y la asignación de tratamientos.

Componentes de un arreglo experimental



Principios básicos del diseño experimental

Son tres los principios del Diseño Experimental:

(a) Replicación

(b) Aleatorización

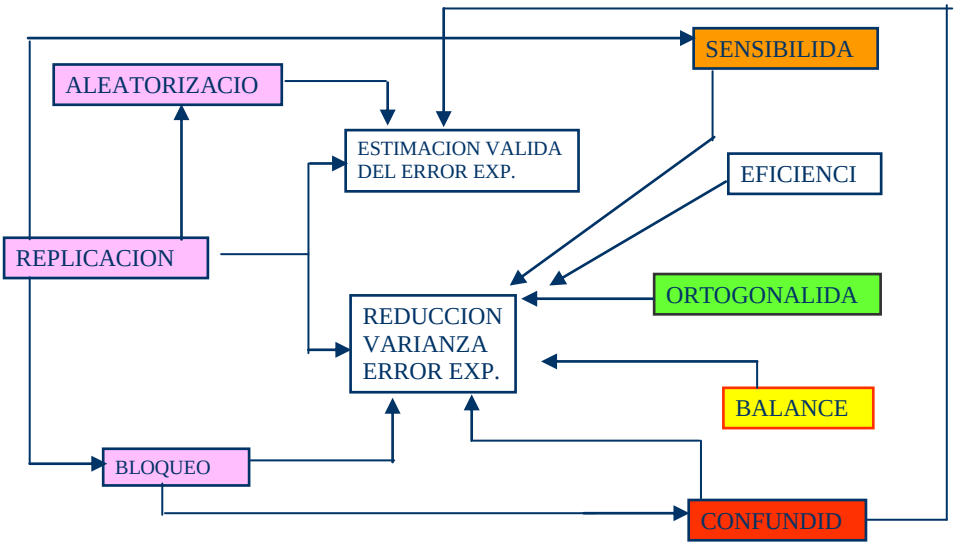
(c) Control Local

Estos tres principios se pueden esquematizar de la siguiente manera:

(a) La replicación permite la validar la estimación del error experimental conjuntamente con la aleatorización.

(b) La replicación conjuntamente con el bloqueo permite la reducción del error experimental.

(c) Por otro lado, la sensibilidad, la eficiencia, ortogonalidad, balance y confusión permiten la reducción del error experimental.



Modelos de Diseño Experimental

Experimental	Cuasiexperimental	No experimental
Manipulación Control Aleatorización	Manipulación Sin control Sin Aleatorización	Sin Manipulación Sin control Sin aleatorización
DCA DBCA DCL	Diseños de series de tiempo.	Correlacional (o ex post facto) Descriptiva

Fases del diseño de experimentos

Reconocimiento y formulación del problema.

Selección de los factores y sus niveles.

Selección de las variables respuestas.

Selección del Diseño Experimental.

Realización del Experimento.

Análisis de datos.

Conclusiones y Recomendaciones.

Definición de Términos Fundamentales

Tratamiento: es una de las modalidades o niveles que puede tomar un factor en estudio, por ejemplo, las variedades por ensayar en un experimento de campo.

Unidad Experimental: es la unidad básica más elemental que se emplea para experimentar; sobre las unidades se experimentales se aplican los tratamientos que son objeto de investigación. En la experimentación agrícola se denominan simplemente parcelas.

Bloque: es un conjunto de unidades experimentales más o menos homogéneas, cuyo objetivo es eliminar la variabilidad del material experimental.

Repetición: es el número de veces que se repite el experimento básico (aplicar los tratamientos a las unidades aleatoriamente). En la experimentación agrícola, es muy frecuente denominar con este término a los bloques de un experimento.

Aleatorización: es el proceso de asignación de los tratamientos a las unidades experimentales, sin que intervenga para ello la voluntad del investigador.

El proceso se realiza usando la tabla de números aleatorios, tarjetas numeradas, o una calculadora de bolsillo, o un paquete de software que genere números aleatorios.

Balanceo: es la asignación de igual número de unidades experimentales por tratamiento. Su uso aumenta la precisión del ensayo.

Control Local: es el conjunto de estrategias que permiten lograr la reducción de la varianza del error experimental.

Entre estas estrategias figuran:

el agrupamiento, el bloqueo, el anidamiento, balanceo, apareamiento, compensación, mediciones repetidas, grupos aleatorios y homogéneos, uso de variables concomitantes o covariables y la confusión.

Agrupamiento: es una estrategia para reducir el error experimental que consiste en colocar unidades experimentales homogéneas dentro de cada grupo.

Anidamiento o Jerarquización: es una estrategia para reducir el error experimental que consiste en colocar dentro de un tratamiento o combinación de tratamiento las unidades experimentales a las cuales interesa medir el efecto de tratamiento. De la definición se desprende que estas unidades experimentales no aparecerán en aquella combinación de tratamiento cuyo efecto no se desea medir.

Sensibilidad: es la capacidad que tiene el diseño experimental de detectar pequeñas diferencias entre los tratamientos, por pequeñas que estas sean.

Ortogonalidad: es la independencia entre tratamientos o efectos de tratamientos.

Eficiencia: es la comparación entre las varianzas del error de dos o más diseños experimentales, para expresar mediante este criterio cual de los dos es más eficiente.

Confundido: es una estrategia para reducir el error experimental y permitir el menor uso de unidades experimentales en algunos diseños. Consiste en confundir unos factores con otros para utilizar menor número de unidades experimentales.

Aditivo Lineal: es una expresión algebraica que resume todos los componentes que están participando en un experimento: tratamientos, bloques, efectos confundidos, efectos aleatorios y otros. Permite la expresión de la naturaleza de los efectos y las restricciones del modelo matemático.

Variable Respuesta: es la medición de la respuesta de los tratamientos conocida como variable dependiente.

Variable independiente: es el conjunto de factores cuyo efecto sobre la variable dependiente se desea medir.

Capítulo II

Diseño Completamente Aleatorizado

Es el diseño más simple y sencillo de realizar, en el cual los tratamientos se asignan al azar entre las unidades experimentales (UE) o viceversa. Este diseño tiene amplia aplicación cuando las unidades experimentales son muy homogéneas, es decir, la mayoría de los factores actúan por igual entre las unidades experimentales. Esta situación se presenta en los experimentos de laboratorio donde casi todos los factores están controlados. También en ensayos clínicos y en experimentos industriales. En ensayos de invernaderos es muy útil. Ha sido ampliamente utilizado en experimentos agrícolas.

La homogeneidad de las unidades experimentales puede lograrse ejerciendo un control local apropiado (seleccionando, por ejemplo, sujetos, animales o plantas de una misma edad, raza, variedad o especie). Pero debe tenerse presente que todo material biológico, por homogéneo que sea, presenta una cierta fluctuación cuyos factores no se conocen y son, por lo tanto, incontrolables.

En este mismo orden de ideas, si logramos controlar factores cualitativos como sexo, camada, color, raza o cuantitativos como peso, alzada, edad, consumo, podremos eliminar su influencia del error experimental; la varianza de éste componente disminuiría y, en consecuencia, aumentaría la eficiencia del experimento posibilitando la detección de efectos entre los tratamiento o condiciones experimentales si es que los hay.

Su nombre deriva del hecho que existe completamente una aleatorización, la cual valida como ya se dijo la prueba F de Fisher-Snedecor. También se le conoce como Diseño de una Vía o un sólo criterio de clasificación en virtud de que las respuestas se hallan clasificadas únicamente por los tratamientos.

Este diseño no impone ninguna restricción en cuanto a las unidades experimentales, éstas deberán ser, en todo caso, homogéneas.

El diseño en su estructura no se ve afectado por el número igual o desigual de observaciones por tratamiento.

Modelo Aditivo Lineal

El modelo aditivo lineal es una expresión algebraica que condensa todos los factores presentes en la investigación. Resulta útil para sintetizar que factores son independientes o dependientes, cuáles son fijos o aleatorios, cuáles son cruzados o anidados.

Para este diseño el modelo aditivo lineal es:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

Donde:

Y_{ij} : es la respuesta (variable de interés o variable medida)

μ : es la media general del experimento

τ_i : es el efecto de tratamiento

ε_{ij} : es el error aleatorio asociado a la respuesta Y_{ij} .

Modelo de efectos fijos, aleatorios y mixtos

Como se observa el efecto de tratamiento puede ser fijo o aleatorio. En otros casos Mixto.

Modelo I (o Modelo de Efectos Fijos)

Cuando los factores son fijos el investigador ha escogido los factores en forma no aleatoria y sólo está interesado en ellos.

En este caso el investigador asume que $\sum \tau_i = 0$, lo cual refleja la decisión del investigador de que únicamente está interesado en los t tratamientos presentes en el experimento. La mayor parte de los experimentos de investigación comparativa pertenecen a este modelo.

Modelo II (Modelo de efectos aleatorios o modelo de componentes de varianza)

Cuando los factores son aleatorios, el investigador, selecciona al azar los de interés de varios que dispone y los asigna a las unidades experimentales.

En este caso el investigador asume que los τ_i tratamientos están distribuidos normal e independientemente con media cero y varianza sigma cuadrado, lo cual se acostumbra a abreviar así: DNI $(0, \sigma_\tau^2)$, lo cual refleja la decisión del investigador de que sólo está interesado en una población de tratamientos, de los cuales únicamente una muestra al azar (los t tratamientos) están presentes en el experimento.

Modelo Mixto

Hace referencia a aquellos casos en los cuales el investigador considera tanto factores fijos como aleatorios en el mismo experimento.

La especificación del modelo es importante en el diseño y análisis del experimento. Su importancia radica en el hecho de que sirven para indicar cuales van a ser los cuadrados medios esperados. Estos cuadrados medios esperados se utilizan para derivar las pruebas exactas de F.

A manera de ejemplo se exponen los cuadrados medios esperados para igual y desigual número de observaciones por tratamiento

Igual Número de Observaciones por Tratamiento		
Fuente de Variación	Modelo I	Modelo II
TRATAMIENTO	$\sigma^2 + n\sum \tau_i^2/(t-1)$	$\sigma^2 + n\sigma_\tau^2$
ERROR EXP.	σ^2	σ^2

Desigual Número de Observaciones por Tratamiento		
Fuente de Variación	Modelo I	Modelo II
TRATAMIENTO	$\sigma^2 + \sum n_i \tau_i^2/(t-1)$	$\sigma^2 + n_0\sigma_\tau^2$
ERROR EXP.	σ^2	σ^2

Más adelante se expondrán las reglas que permiten obtener estos cuadrados medios esperados; por los momentos, aceptemos los resultados como se presentan.

Las pruebas exactas de F se pueden obtener con los cuadrados medios esperados. Por ejemplo, la prueba exacta de F para tratamientos para el modelo II con igual número de observaciones es:

$$F = \frac{\sigma^2 + n\sigma_\tau^2}{\sigma^2}$$

Según este cociente de varianzas, F deberá tener un valor cercano a uno, cuando las medias de tratamientos sean iguales, esto es $\left(\sigma_\tau^2 = 0\right)$ y deberá aumentar cuando las medias de tratamientos μ_i difieran de manera considerable.

Representación Simbólica para el Diseño Completamente Aleatorizado.

La representación esquemática para este diseño es:

Tratamientos

	1	2	…	k
Observaciones	Y11 Y12 Y13 . . . Y1n1	Y21 Y22 Y23 . . . Y2n2	… … … … … … …	Yk1 Yk2 Yk3 . . . Yknk
Totales	Y1.	Y2.	…	Yk.
Medias	$\overline{Y}_{1.}$	$\overline{Y}_{2.}$	…	$\overline{Y}_{k.}$

La representación esquemática plantea la forma como se van a determinar los efectos:

(i) El Efecto de tratamiento (τ_i) viene dado por:

$$\tau_i = \sum\sum(\overline{Y}_{i.} - \overline{Y}_{..})^2$$

Esta suma de cuadrados de tratamiento puede ser calculada así:

SCTRAT= $\sum n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2$

O también:

SCTRAT= $\sum \frac{Y_{i.}^2}{n_i} - FC$

Donde $FC = \frac{(\sum\sum Y_{ij})^2}{N}$

(ii) El efecto de la variación total puede ser determinado por:

SCTOTAL= $\sum\sum (\bar{Y}_{ij} - \bar{Y}_{..})^2$

O también:

SCTOTAL= $\sum\sum Y_{ij}^2 - FC$

(iii) El efecto del error experimental puede ser estimado por la suma de cuadrados totales menos la suma de cuadrados de los tratamientos, esto es:

SCERROR=SCT - SCTRAT

Cuadro de anova

El cuadro de análisis de varianza (anova) es un arreglo dado por las fuentes de variación, seguido de los grados de libertad, de las sumas de cuadrados, de los cuadrados medios de cada componente, así como del valor F y su probabilidad de significación (valor P).

El esquema es como sigue:

Fuentes de Variación	Grados de libertad	Suma de Cuadrados	Cuadrados Medios	Valor F	Valor P(*)
TRAT	k-1	SCTRAT	$\frac{SCTRAT}{k-1} = CMTRAT$	$\frac{CMTRAT}{CMERROR}$	
ERROR	$\sum(n_i - 1)$	SCERROR	$\frac{SCERROR}{\sum(n_i - 1)} = CMERROR$		
TOTAL	$\sum n_i - 1$	SCT	$\frac{SCT}{\sum n_i - 1} = CMTOTAL$		

(*) El valor P suele ser fijado arbitrariamente en: 0,05; 0,01 ó 0,001.

Diseño completamente al azar con submuestreo

En algunas ocasiones es necesario trabajar con unidades experimentales muy grandes, digamos por ejemplo, parcelas de gran tamaño o toda una hacienda o finca como unidad. También puede necesitarse realizar algunas determinaciones en los experimentos las cuales serían muy tediosas tomarlas en toda la unidad experimental, siendo por esto necesario extraer subunidades de cada unidad experimental y esta situación genera un submuestreo de observaciones por unidad experimental. Este diseño completo al azar de submuestreo da origen a muestras anidadas dentro de otras, ocasionando una variante del diseño.

Las muestras anidadas dentro de otras permiten establecer al mismo tiempo una estrategia de control local referida a los niveles de tratamiento porque sólo ciertos niveles de factores se probarán en algunas combinaciones de tratamiento y en otras no.

Veamos un ejemplo, un experimentador puede estar interesado en probar ciertas raciones (R1, R2,…,Rk) en n animales pero para cada animal se realizaron dos determinaciones $(n_{ij} = 2)$ de los porcentajes de grasa de la leche. Los datos aparecen a continuación:

R1	R2	…	Rk
Y111 Y112	Y211 Y212	… …	Yk11 Yk12
Y121 Y122	Y221 Y222	… …	Yk21 Yk22
…	…	…	…
Y1n1 Y1n2	Y2n1 Y2n2	… …	Ykn1 Ykn2

En este caso particular existen dos fuentes de variabilidad que contribuyen a formar la varianza para las comparaciones entre los promedios de los tratamientos, estas son:

(i) El error de muestreo que es la variación entre subunidades de una misma unidad experimental. Y se puede estimar su valor numérico a través del cuadrado medio del error de muestreo.

(ii) El error experimental que es la variación entre unidades experimentales sometidas a un mismo tratamiento.

Modelo aditivo lineal

El modelo aditivo lineal entonces deberá ser reformulado para incluir el término correspondiente al error de muestreo:

$$Y_{ijk} = \mu + \tau_i + \varepsilon_{ij} + \delta_{ijk}$$

Donde:

I=1,2,3,…,k; J=1,2,3,…,ni y l=1,2,3,…,nij

μ = es el verdadero efecto medio

τ_i : es el efecto del i-ésimo tratamiento

ε_{ij} : es el efecto de la j-ésima unidad experimental sujeta al i-ésimo tratamiento (o error experimental).

δ_{ijl} : es el efecto de la l-ésima observación tomada de la j-ésima unidad experimental sujeta al i-ésimo tratamiento (o error de muestreo).

Estimación de los efectos en el caso de submuestreo

$\overline{Y}_{...}$: es un estimador de μ

$(\overline{Y}_{i.} - \overline{Y}_{...})$: es el estimador de los efectos de tratamiento τ_i

$(\overline{Y}_{ij.} - \overline{Y}_{i.})$: es el estimador del error experimental ε_{ij}

$(Y_{ijl} - \overline{Y}_{ij.})$: es el estimador del error de muestreo δ_{ijl}

Hipótesis Estadísticas: Nula y Alterna

$$Ho: \mu_1 = \mu_2 = \mu_3 = ... = \mu_k$$

$$H_1: \mu_1 \neq \mu_2 \neq \mu_3 \neq ... \neq \mu_k$$

Es decir,

Ho: no existe diferencia significativa entre los tratamientos

H1: existe diferencia significativa entre los tratamientos

Cuadro de ANOVA en el caso de submuestreo

Fuentes de Variación	Grados de Libertad	Suma de Cuadrados	Cuadrados Medios	F
TRAT	K-1	SCTRAT	CMTRAT	$CMTRAT \div CMERROR\_EXP$
ERROR EXP:	$\sum(n_1 - 1)$	SCERROR_EXP	CMERROR EXP	
ERROR MUESTREO	$\sum\sum(n_{ij} - 1)$	SCERROR_ MUESTREO	CMERROR MUESTREO	
TOTAL	N-1	SCT		

Sumas de Cuadrados en el caso de ANOVA con Submuestreo

Las sumas de cuadrados que aparecen en el cuadro pueden calcularse así:

Suma de Cuadrados Totales:

SCT= $\sum\sum\sum Y_{ijl}^2 - FC$

Donde $FC = \frac{Y_{...}^2}{n_i n_{ij} k}$

Suma de Cuadrados de tratamientos:

$$SCTRAT = \sum \frac{\left(\sum\sum Y_{ijl}\right)^2}{\sum n_{ij}} - FC$$

Suma de Cuadrados del Error Experimental:

$$\text{SCEE}=\sum\sum\frac{\left(\sum Yijl\right)^2}{\sum n_{ij}}-\sum\frac{\left(\sum\sum Y_{ijl}\right)^2}{\sum n_{ij}}$$

Suma de Cuadrados del Error de Muestreo:

$$\text{SCEM}=\left[\sum\sum\sum Y_{ijl}^2-\frac{\left(\sum\sum\sum Y_{ijl}\right)^2}{N}\right]-\left(SCTRAT+SCERROR\_EXP\right)$$

Donde:

$$\text{N}=\sum\sum n_{ij}$$

Estas fórmulas de cálculo hacen referencia al caso general y por ende de desigual número de observaciones por unidad experimental. Pero en caso de existir igual número de observaciones por unidad experimental las fórmulas anteriores se simplifican grandemente:

La Suma de Cuadrados Totales:

$$\text{SCT}=\sum\sum\sum Y_{ijl}^2-FC$$

La Suma de Cuadrados de Tratamientos:

$$\text{SCTRAT}=nm\sum\left(\bar{Y}_i-\bar{Y}_{...}\right)^2$$

La Suma de Cuadrados del Error Experimental:

$$\text{SCERROR\_EXP}=m\sum\sum\left(\bar{Y}_{ij}-\bar{Y}_{i..}\right)^2$$

La Suma de Cuadrados del Error de Muestreo:

SCERROR_MUESTREO=

$$\left[\sum\sum\sum Y_{ijl}^2-\frac{\left(\sum\sum\sum Y_{ijl}\right)^2}{tnm}\right]-\left(SCTRAT+SCERROR\_EXP\right)$$

Proceso de aleatorización en el diseño completamente aleatorizado

El proceso de distribución de los tratamientos al azar en las unidades experimentales se puede realizar usando una tabla de números aleatorios o mediante un algoritmo computarizado de SAS.

Se utilizará el primero de éstos por ser de uso más frecuente.

Supongamos un experimento donde deseamos probar 4 tipos diferentes de hormonas A,B,C y D, cada una en una dosis única, para determinar su efecto sobre la capacidad de aumento de peso en ovejas. Se desean realizar 5 repeticiones.

Se procede así:

- Se forman grupos homogéneos en cuanto a una variable (digamos en este caso peso)
- Cada grupo va a contener 4 ovejas.
- Realizando el sorteo, mediante la tabla de números aleatorios, puede resultar así:

1	A	B	C	D
2	D	C	B	A
3	B	C	A	D
4	C	B	D	A
5	B	A	D	C

De esta forma, quedan distribuidos los tratamientos entre las unidades experimentales, que en total son: 5X4=20 ovejas.

El balance existe en este caso cuando permitimos que cada repetición (replicación) contenga todos los tratamientos.

Ventajas del diseño completamente al azar

(1) Su flexibilidad. Puesto que permite una total libertad en el dispositivo experimental, por un lado, se puede probar cualquier número de tratamiento y, por otro lado, el número de observaciones por tratamiento puede ser igual o desigual.

(2) Maximiza los grados de libertad para estimar el error experimental.

(3) Permite observaciones perdidas y no se dificulta el análisis estadístico.

(4) Muy fácil de usar en experimentación.

Desventajas del diseño completamente al azar

(1) Es apropiado para un número pequeño de tratamientos.

(2) Es apropiado sólo en caso de disponer de material experimental homogéneo.

(3) En comparación con otros dispositivos experimentales donde se pueda ejercer control local es menos sensible y tiene un poder analítico débil. Esta falta de sensibilidad se debe a que todos los factores que no se controlaron se acumulan en el error experimental, provocando que éste aumente y no permita detectar diferencias significativas entre los tratamientos. Este es el motivo por el cual los investigadores estadígrafos se dedicaron a buscar diseños experimentales más complejos, pero que fueran capaz de estimar el error experimental con mayor grado de precisión.

Supuestos del Anova completamente al azar

Son cuatro los supuestos que debe cumplir el ANOVA, que se pueden abreviar con la nemotécnia: HINA. La H de homogeneidad de las varianzas, I de independencia de los errores, la N de normalidad de los efectos y la A de aditividad de los efectos:

- **Homogeneidad de las varianzas** se refiere a que cada respuesta Y_{ij} debe poseer dentro de cada tratamiento una variación parecida o igual a la de otro tratamiento. Este supuesto puede ser probado postulando como hipótesis nula y alterna las siguientes:

  $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots = \sigma_k^2$ (los k tratamientos tienen igual varianza)

  $H_1 : \sigma_1^2 \neq \sigma_2^2 \neq \sigma_3^2 = \ldots \neq \sigma_k^2$ (no todos los k tratamientos tienen igual varianza)

Para verificar el cumplimiento de este supuesto se utiliza la prueba de Bartlett o de Levene. La aplicación de estas pruebas debe conducir a un no rechazo de Ho para que exista la homogeneidad de varianzas.

- **Independencia de los errores de un tratamiento a otro:** Es decir, se calculan los errores residuales para cada grupo de tratamiento a través de la fórmula:

  $$ERROR\_RESIDUAL = Y_i - \overline{Y}_i$$

  Las hipótesis son:

  Ho: Hay independencia de errores

  H1: Hay dependencia de errores

  El supuesto de independencia se verifica su cumplimiento utilizando la prueba de aleatoriedad o de las rachas que es una prueba del ámbito no paramétrico.

- **Normalidad de los efectos:** se refiere a que las respuestas Y_{ij} deben poseer una distribución normal dentro de cada grupo de tratamiento. Para verificar su cumplimiento se plantean las hipótesis:

  Ho: La distribución de cada tratamiento es normal

  H1: La distribución de cada tratamiento no es normal

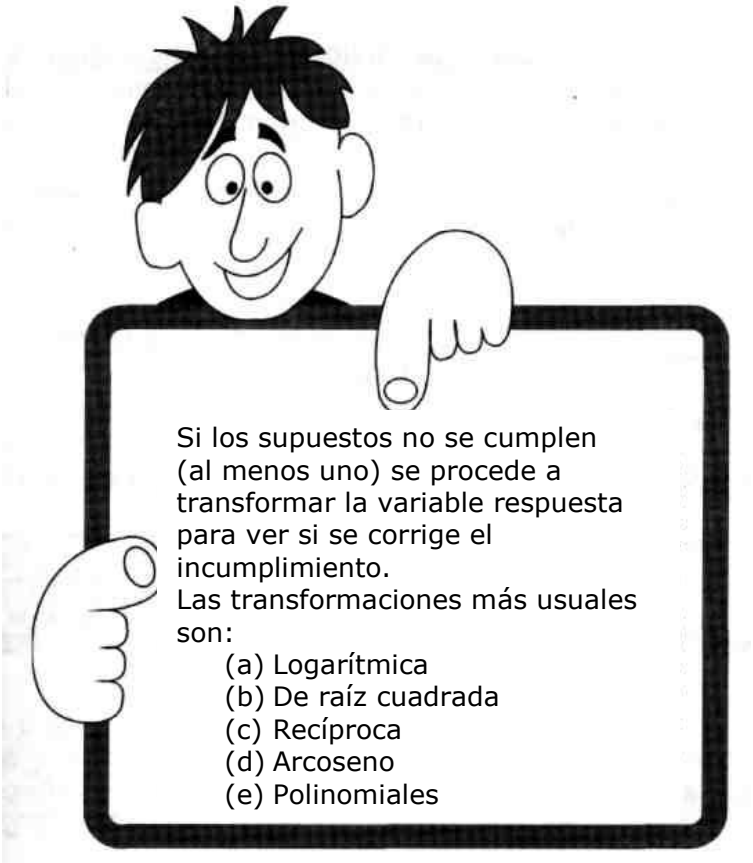
  La prueba que utilizamos en este caso es la de Shapiro-Wilk para muestras pequeñas y Kolmogorov-Smirnov para muestras grandes. Aunque la Prueba de Shapiro-Francia es equivalente a la de Kolmogorov-Smirnov.

- **Aditividad de los efectos:** las respuestas de cada grupo de tratamiento es la suma de la media general más un efecto aleatorio asociado a la respuesta. El sistema hipotético es:

  Ho: Los efectos son aditivos

  H1: Los efectos no son aditivos

La prueba estadística apropiada es la Prueba de Tukey para la aditividad. Esta prueba debe conducir a un no rechazo de Ho para que el supuesto se cumpla.



Ejemplo numérico del diseño completamente al azar

Se desea probar la hipótesis de que las notas de estadísticas en pruebas objetivas cortas dependen de la hora de realización de la prueba. Para ello se han escogido, al azar, cinco alumnos del turno matutino, vespertino y nocturno. Las pruebas arrojaron los resultados siguientes:

Matutino	Vespertino	Nocturno
16	10	15
17	11	08
18	12	09
19	13	13
20	14	14

Se pide:

(a) Plantear hipótesis nula y alterna

(b) Escribir la matriz de datos a ser procesada

(c) Establecer la secuencia a seguir en el STATISTIX

(d) Exponer la salida del software

(e) De existir diferencia significativa entre los turnos, verificar, ¿ cuál es el mejor?

(f) Concluir

Solución:

(a) Hipótesis Estadísticas: Nula y Alterna

$$Ho: \mu_1 = \mu_2 = \mu_3 = ... = \mu_k$$

$$H_1: \mu_1 \neq \mu_2 \neq \mu_3 \neq ... \neq \mu_k$$

Es decir,

Ho: no existe diferencia significativa entre los tratamientos

H1: existe diferencia significativa entre los tratamientos

Respecto al problema planteado:

Ho: los rendimientos no difieren por turno

H1: los rendimientos difieren por turno

(b) La matriz de datos es:

Codificando los turnos como 1=matutino, 2=vespertino y 3=nocturno, tenemos:

TURNO	NOTA
1	16
1	17
1	18
1	19
1	20
2	10
2	11
2	12
2	13
2	14
3	15
3	08
3	09
3	13
3	14

(C) Secuencia en el Statistix

Para la entrada de datos:

DATA>INSERT>VARIABLE

NOMBRES DE LAS VARIABLES

TIPEAR: TURNO NOTA OPRIMIR OKEY

VACIAR LOS DATOS

(DESPUÉS DE CADA DATO PRESIONAR ENTER)

GUARDAR ARCHIVO: REND.sx

Para el procesamiento estadístico:

STATISTICS

ONE, TWO, MULTIPLE SAMPLE TESTS

ONE WAY AOV

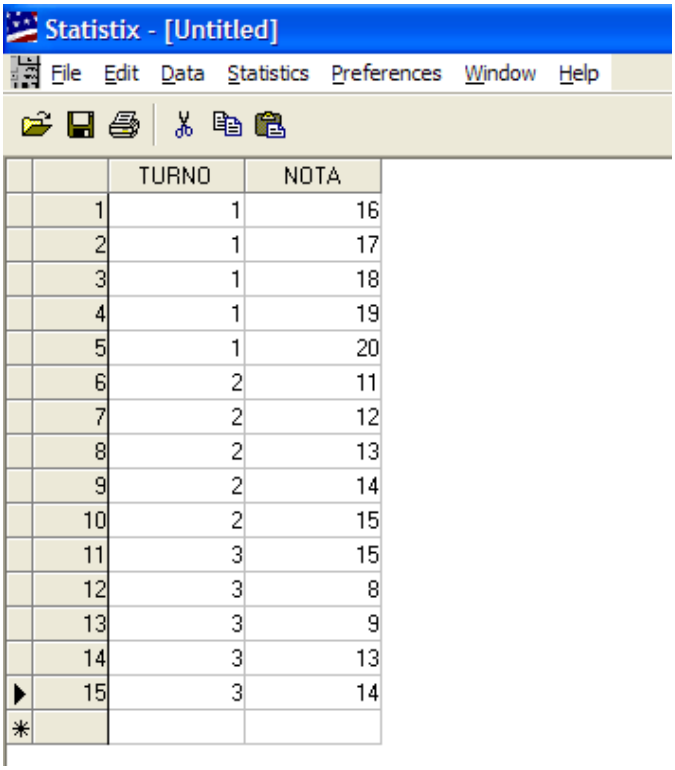
EN LA CAJA DE DIÁLOGO ESCRIBIMOS

NOTA (COMO VARIABLE DEPENDIENTE)

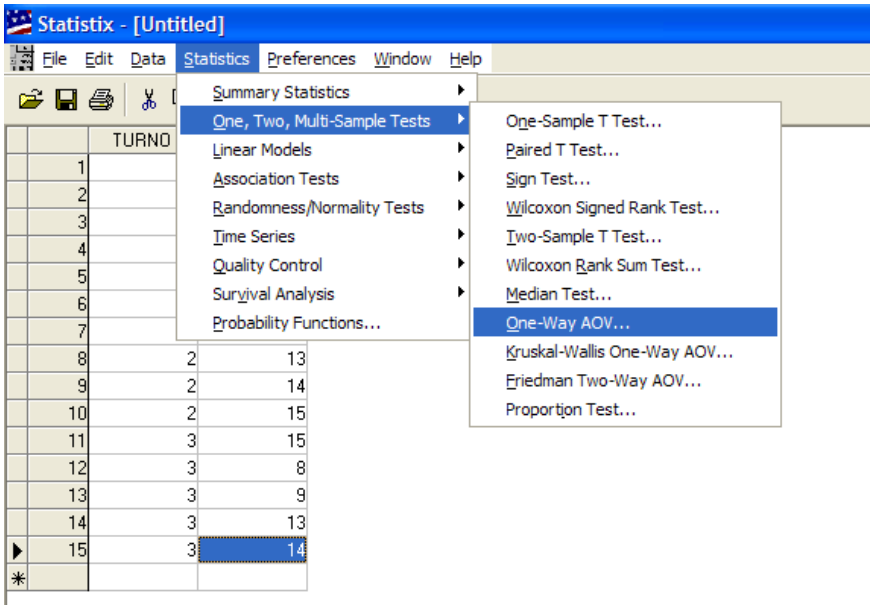
TURNO (COMO VARIABLE CATEGÓRICA)

RESULTADOS (PARA REALIZAR COMPARACIONES MÚLTIPLES)

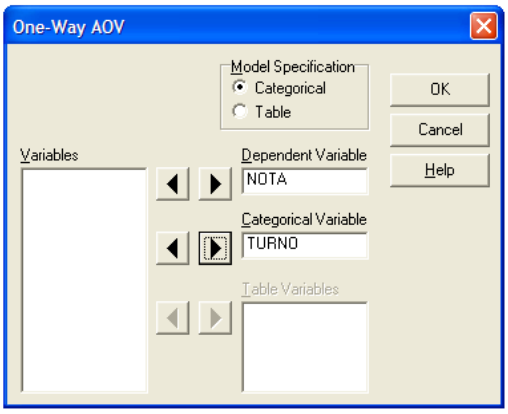
PRUEBA TUKEY



	TURNO	NOTA
1	1	16
2	1	17
3	1	18
4	1	19
5	1	20
6	2	11
7	2	12
8	2	13
9	2	14
10	2	15
11	3	15
12	3	8
13	3	9
14	3	13
15	3	14



	TURNO	
1		
2		
3		
4		
5		
6		
7		
8	2	13
9	2	14
10	2	15
11	3	15
12	3	8
13	3	9
14	3	13
15	3	14

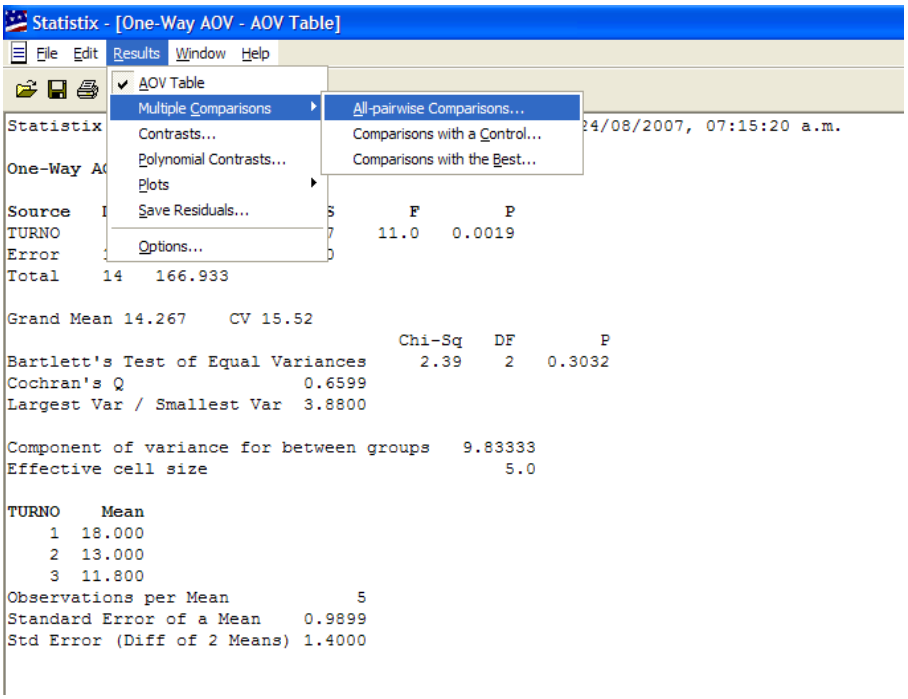


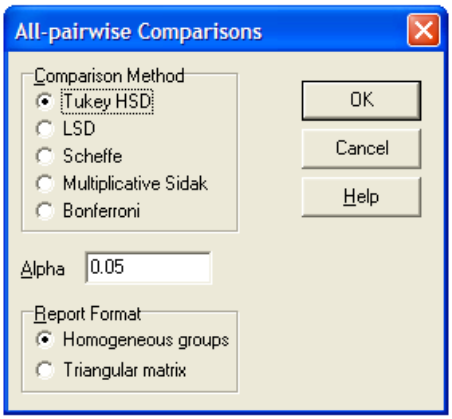
SALIDA DEL SOFTWARE:

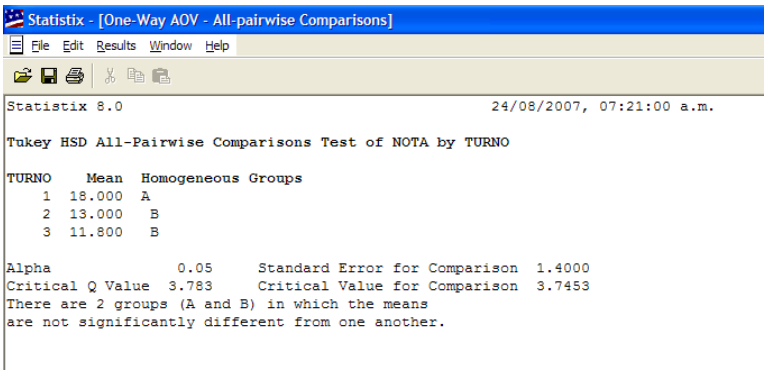
Statistix - [One-Way AOV - AOV Table]

File Edit Results Window Help

```
Statistix 8.0                                        24/08/2007, 07:15:20 a.m.

One-Way AOV for NOTA by TURNO

Source    DF         SS         MS         F         P
TURNO      2    108.133    54.0667    11.0    0.0019
Error     12     58.800     4.9000
Total     14    166.933

Grand Mean 14.267    CV 15.52
                                        Chi-Sq    DF         P
Bartlett's Test of Equal Variances        2.39     2    0.3032
Cochran's Q                   0.6599
Largest Var / Smallest Var    3.8800

Component of variance for between groups    9.83333
Effective cell size                             5.0

TURNO     Mean
    1   18.000
    2   13.000
    3   11.800
Observations per Mean             5
Standard Error of a Mean     0.9899
Std Error (Diff of 2 Means)  1.4000
```


Statistix - [One-Way AOV - AOV Table]
File Edit Results Window Help
AOV Table
Multiple Comparisons
Contrasts...
Polynomial Contrasts...
Plots
Save Residuals...
Options...
All-pairwise Comparisons...
Comparisons with a Control...
Comparisons with the Best...
Statistix
24/08/2007, 07:15:20 a.m.
One-Way AO
Source I S F P
TURNO 7 11.0 0.0019
Error 0
Total 14 166.933
Grand Mean 14.267 CV 15.52
Chi-Sq DF P
Bartlett's Test of Equal Variances 2.39 2 0.3032
Cochran's Q 0.6599
Largest Var / Smallest Var 3.8800
Component of variance for between groups 9.83333
Effective cell size 5.0
TURNO Mean
1 18.000
2 13.000
3 11.800
Observations per Mean 5
Standard Error of a Mean 0.9899
Std Error (Diff of 2 Means) 1.4000


All-pairwise Comparisons
Comparison Method
Tukey HSD
LSD
Scheffe
Multiplicative Sidak
Bonferroni
OK
Cancel
Help
Alpha 0.05
Report Format
Homogeneous groups
Triangular matrix


Statistix - [One-Way AOV - All-pairwise Comparisons]
File Edit Results Window Help
Statistix 8.0
24/08/2007, 07:21:00 a.m.
Tukey HSD All-Pairwise Comparisons Test of NOTA by TURNO
TURNO Mean Homogeneous Groups
1 18.000 A
2 13.000 B
3 11.800 B
Alpha 0.05 Standard Error for Comparison 1.4000
Critical Q Value 3.783 Critical Value for Comparison 3.7453
There are 2 groups (A and B) in which the means
are not significantly different from one another.

Conclusiones:

(a) La prueba F del análisis global del ANOVA indica una diferencia altamente significativa entre los rendimientos por turno (F=11,0;P=0,0019)

(b) La prueba de Tukey indica que el turno de mejor rendimiento es el matutino con un Rendimiento=18,0

(c) Los turnos vespertino y nocturno no difieren en rendimiento.

Capítulo III

Diseños de bloques completos al azar

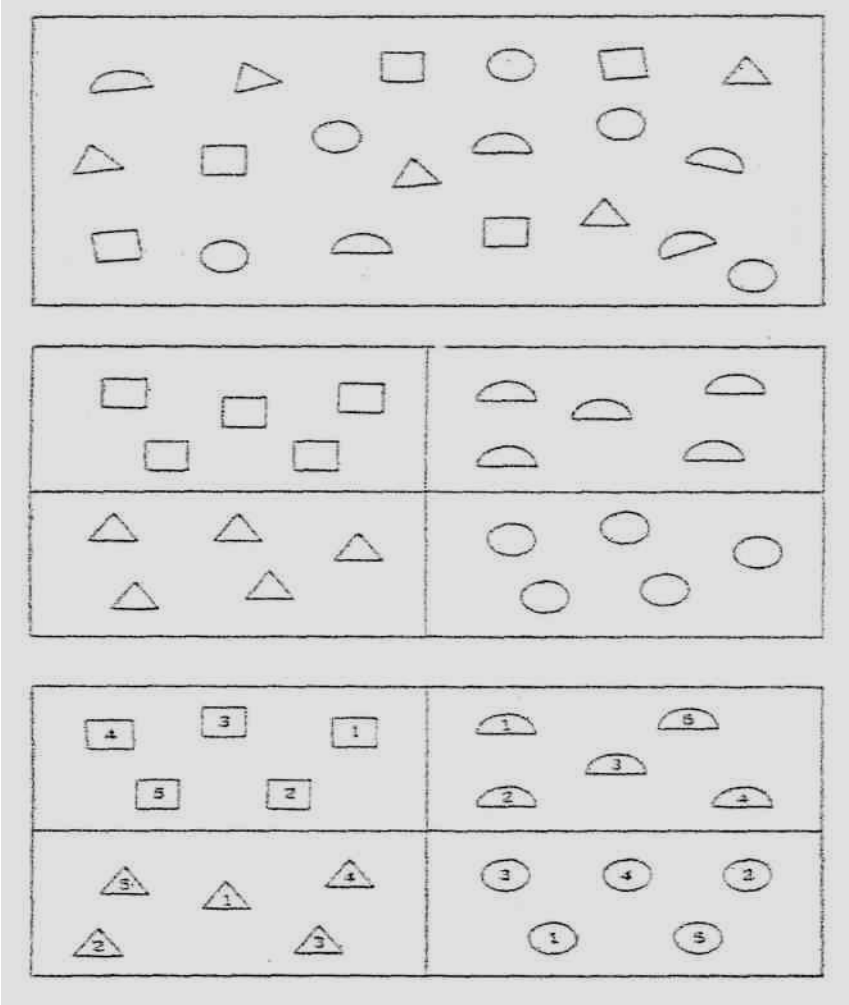
El diseño de bloques completos al azar surge por la necesidad que tiene el investigador de ejercer un control local de la variación dado la existencia de un material experimental heterogéneo.

En ese orden de ideas, los pasos que el investigador sigue son:

(a) Forma los bloques de unidades experimentales homogéneos fundamentándose para ello en algún criterio de bloqueo o agrupamiento. Estos criterios pueden ser: Raza, Época, Edad, Sexo, Peso, Sistema de Manejo, Tipo de Explotación, Zona, País, Número de Partos, número de lactaciones, número de ordeños, corrales o establos, potreros, camadas, métodos, variedades, entre otros.

(b) Luego de formados los bloques se asignan al azar los tratamientos a la unidades experimentales de cada bloque.

El esquema dado a continuación ayuda a comprender la filosofía de la formación de bloques.



Los bloques se definen como un conjunto de unidades experimentales homogéneas dentro de sí y heterogéneos entre sí. En los bloques están representados todos los tratamientos.

Objetivos

(a) Maximizar las diferencias entre bloques

(b) Minimizar la variación dentro de bloques

Ventajas

(a) Elimina una fuente de variación del error, aumentando de esta forma la precisión del ensayo. La precisión del ensayo se mide a través del coeficiente de variación $CV(\%) = \frac{S}{\bar{Y}} x100$

(b) Permite una gran flexibilidad en la relación tratamiento bloque, siempre y cuando se reserven un número igual (o un múltiplo) de tratamientos por unidad experimental.

(c) La pérdida de información por bloque o tratamiento no dificulta el análisis estadístico.

(d) Permite aplicar el principio de confusión al hacer coincidir los bloques con los ciertas variables que influyen sobre la respuesta pero que no son de interés para el investigador.

Desventajas

(a) No es apropiado para un número elevado de tratamientos (se recomienda entre 6 y 24 tratamientos)

(b) No es aconsejable cuando exista una gran variación (en más de una variable) en el material experimental.

(c) Si el efecto de bloque no es significativo se trabaja innecesariamente con disminución de los grados de libertad para el error y la consecuente disminución de la precisión.

(d) Si resulta una interacción entre bloque y tratamiento, se invalida la prueba F.

Modelo aditivo lineal

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

De izquierda a derecha tenemos la variable respuesta Y_{ij}, la media general μ, el efecto del i-ésimo tratamiento τ_i, el efecto de bloque β_j y el efecto del error experimental ε_{ij}.

Representación Esquemática del Diseño en Bloques Completos al Azar

Tratamientos

Bloques	1…	J…	t	Total	Media Bloque
1	Y11	Y1J	Y1t	Y1.	$\overline{Y}_{1.}$
I	YI1	Yij	Yit	Yi.	$\overline{Y}_{i.}$
r	Yr1	Yrj	Yrt	Yr.	$\overline{Y}_{r.}$
Total	$Y_{.1}$	$Y_{.j}$	$Y_{.t}$	$Y_{..}$	
Media TRAT	$\overline{Y}_{.1}$	$\overline{Y}_{.j}$	$\overline{Y}_{.t}$		$\overline{Y}_{..}$

Las fórmulas de definición y de cálculo son:

Factor de Corrección:

$$FC = \frac{Y_{..}^2}{rt}$$

Suma de Cuadrados Totales:

$$\text{SCT} = \sum\sum\left(Y_{ij} - \overline{Y}_{..}\right)^2 = \sum\sum Y_{ij}^2 - FC$$

Cuadro de ANOVA

Fuentes de Variación	Grados de libertad	Suma de Cuadrados	Cuadrados Medios	Valor F
TRAT BLOQ ERROR	t-1 r-1 (t-1)(r-1)	SCTRAT SCBLOQ SCERROR	CMTRAT CMBLOQ CMERROR	$CMTRAT \div CMERROR$
Total	rt-1	$\sum\sum\left(Y_{ij} - \bar{Y}_{..}\right)^2$		

Suma de cuadrados de tratamiento:

$$\text{SCTRAT} = r\sum(\bar{Y}_{.j} - \bar{Y}_{..})^2 = \frac{\sum\sum Y_{ij}^2}{r} - FC$$

Suma de cuadrados de bloques:

$$\text{SCBLOQ} = t\sum(\bar{Y}_{i.} - \bar{Y}_{..})^2 = \frac{\sum Y_{i.}^2}{t} - FC$$

Suma de cuadrados del error:

$$\text{SCERROR} = \sum\sum(Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2 = \text{SCT-(SCTRAT+SCBLOQ)}$$

Diseños en Bloques al Azar con más de una observación por Unidad Experimental o sub-muestreo en Bloques

Tratamientos

BLOQ	1	2	…	t	totales	medias
R1	Y111 . . . Y11S	Y211 . . . Y21S	… … … …	Yt11 . . . Yt1S		
R2	Y121 . . . Y12S	Y221 . . . Y22S	… … … …	Yt21 . . . Yt2S	$Y_{.j.}$	$\overline{Y}_{.j.}$
Rr	Y1r1 . . . Y1rS	Y2r1 . . . Y2rS	… … … …	Ytr1 . . . YtrS		
Totales		$Y_{i..}$			Y…	
Medias		$\overline{Y}_{i..}$				$\overline{Y}_{...}$

Modelo aditivo lineal

$$Y_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ij} + \delta_{ijk}$$

i=1,…,K

j=1,…,r

k=1,…,s

donde:

μ: es la media general

τ_i : es el efecto de tratamiento

β_j : es el efecto de bloque

ε_{ij} : es el error experimental

δ_{ij} : es el error de muestreo

Sumas de Cuadrados en Diseño de Bloques con Submuestreo

La suma de cuadrados totales:

$$SCT = \sum\sum\sum (Y_{ijk} - \bar{Y}_{...})^2 = \sum\sum\sum Y_{ijk}^2 - \frac{Y_{...}^2}{krs}$$

La suma de cuadrados de tratamientos:

$$SCTRAT = rs\sum (\bar{Y}_{i..} - \bar{Y}_{...})^2 = \frac{\sum Y_{i..}^2}{rs} - \frac{Y_{...}^2}{krs}$$

La suma de cuadrados de bloque:

$$SCBLOQ = ks\sum (\bar{Y}_{.j.} - \bar{Y}_{...})^2 = \frac{\sum Y_{.j.}^2}{ks} - \frac{Y_{...}^2}{krs}$$

La suma de cuadrados del error experimental:

$$SCE = S\sum\sum (\bar{Y}_{ij.} - \bar{Y}_{...})^2 - (SCTRAT + SCBLOQ)$$

O también:

$$SCE = \frac{\sum\sum Y_{ij.}^2}{s} - \frac{Y_{...}^2}{krs}$$

La suma de cuadrados del error de muestreo es:

$$SCEM = \sum\sum\sum (Y_{ijk} - \bar{Y}_{ij.})^2 = \sum\sum (\sum Y_{ijk}^2 - \frac{Y_{ij.}^2}{s})$$

Cuadro de Anova

FV	gl	SC	CM	F
TRAT	k-1	SCTRAT	CMTRAT	CMTRAT/CMEE
BLOQ	r-1	SCBLOQ	CMBLOQ	
ERROR EXP.	(k-1)(r-1)	SCEE	CMEE	
ERROR MUESTREO	kr(s-1)	SCEM	CMEM	
TOTAL	krs-1	SCT		

Observaciones faltantes o perdidas en un diseño de bloques

En los experimentos pueden ocurrir accidentes que dan como resultado la pérdida de una o varias unidades experimentales.

Las observaciones perdidas surgen por varias razones:

(i) Un animal puede destruir las plantas de una o varias parcelas

(ii) Puede ocurrir mortalidad de animales

(iii) Pueden enfermar si ser consecuencia del tratamiento empleado

(iv) Un dato registrado puede estar mal tomado

(v) Errores en la aplicación de un tratamiento

La pérdida de una o varias unidades experimentales anula el *Teorema de la Adición de la Suma de Cuadrados*, y, por consiguiente, no se podría emplear el *método de los mínimos cuadrados,* a menos que se estime un valor para la o las unidades perdidas. Además, las observaciones perdidas destruyen el balance o simetría con la cual fue planificado nuestro experimento originalmente. En tal situación nosotros podemos emplear dos caminos, el análisis estadístico para desigual número de observaciones, digamos por tratamiento, ó, el procedimiento de estimación de observaciones perdidas debido a Yates.

Los casos que se pueden presentar son:

(a) **Falta un bloque o un tratamiento completo:** en este caso se elimina el bloque o el tratamiento y se procede al análisis habitual.

(b) **Falta una observación:** en esta situación se estima la observación perdida por el método de Yates:

$$Y = \frac{tT + rB - G}{(t-1)(r-1)}$$

Donde:

Y: observación perdida, t: número de tratamientos, r: número de bloques, T: suma de observaciones en el tratamiento donde figura la observación perdida, B: suma de observaciones en el bloque donde figura la observación perdida, G: gran total de las observaciones que quedan en el experimento. La observación así estimada se anota en la matriz de datos y se procede al análisis estadístico de la forma habitual, con la excepción de que se reducen en uno los grados de libertad del total y, como consecuencia también, en igual cantidad, los grados de libertad del error.

Yates, indicó que el análisis de varianza desarrollado utilizando valores estimados conlleva a una sobre-estimación de la suma de cuadrados de tratamientos, la cual puede ser corregida a través de la fórmula:

$$SCTRAT\_CORREGIDA = SCTRAT\_SIN\_CORREGIR - \frac{[B-(t-1)Y]^2}{t(t-1)}$$

(c) **Falta más de una observación:** el procedimiento más exacto para esta situación, consiste en asignar aquel valor de la observación perdida que reduzca la suma de cuadrados del error experimental. Esta solución la podemos encontrar en el cálculo diferencial. Calculamos la expresión algebraica de la suma de cuadrados del error experimental, derivamos ésta expresión con respecto a las observaciones perdidas e igualamos a cero, obteniendo de ésta manera un sistema de ecuaciones cuya solución nos da el valor estimado de las observaciones perdidas.

Si el número de observaciones perdidas son dos, ocurre la pérdida de dos grados de libertad para el error y dos para el total. Asimismo, debe reducirse la suma de cuadrados de tratamiento, en igual cantidad, a fin de evitar la sobre-estimación. La fórmula es:

$$SCTRAT\_CORREGIDA = SCTRAT\_SIN\_CORREGIR - \frac{[B'-(t-1)Y1]^2 + [B''-(t-1)Y2]^2}{t(t-1)}$$

Donde:

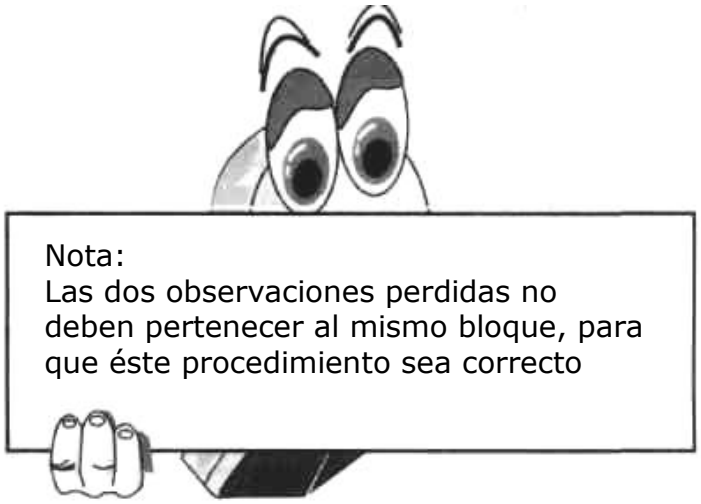
t: es el número de tratamientos

B’: suma de observaciones en el bloque donde figura Y1.

B’’: suma de observaciones en el bloque donde figura Y2.

Y1: estimación de la primera observación perdida

Y2: estimación de la segunda observación perdida



Eficiencia relativa de un diseño

En algunas situaciones es de interés conocer qué tan efectivo fue utilizar un diseño con respecto a otro. Esto es particularmente útil si existen dudas en cuanto a la utilización de un diseño en particular.

Se define como la medida porcentual de la varianza del error de un diseño con respecto a otro. En general, será más eficiente aquel diseño que posea menor varianza del error.

Si deseamos comparar la eficiencia de un diseño dos con respecto al uno, lo indicaremos así:

$$ER(\%) = \frac{S_{\bar{Y}(2)}}{S_{\bar{Y}(1)}} x100$$

Nota: La eficiencia relativa del diseño que se quiere estimar debe colocar en el denominador la varianza del error.

Ejemplo numérico

Si la varianza del error de un diseño en bloques es en un caso particular $S_{\bar{Y}(2)} = 18{,}06$ y la de un diseño completamente al azar es $S_{\bar{Y}(1)} = 7{,}50$. Luego la eficiencia del bloque respecto al diseño completamente al azar es:

$$ER(\%) = \frac{S_{\bar{Y}(2)}}{S_{\bar{Y}(1)}} x100$$

ER(%)= $\frac{18{,}06}{7{,}50} x100 = 240{,}8\%$ (esta es la eficiencia del diseño completamente al azar)

Entonces podemos afirmar que al cambiar de un diseño completamente al azar a un diseño de bloque la ganancia relativa en eficiencia es: 240,8 %- 100%=140,8%

Fisher, sugiere al respecto, que esta medida es poco confiable si no se consideran los grados de libertad del error de los diseños en comparación. Así, la comparación de dos diseños: completamente al azar y el de bloques vendría dado por:

$$\text{Cantidad relativa de Información} = \frac{(n_1+1)(n_2+3)}{(n_2+1)(n_1+3)} x \frac{S^2_{CA}}{S^2_{BA}}$$

Donde:

n1= grados de libertad del error en un diseño en bloques

n2= grados de libertad del error en un diseño completamente al azar

El factor de ajuste de los grados de libertad del error al cual se refiere Fisher es:

$$\text{Factor de Ajuste} = \frac{(n_1+1)(n_2+3)}{(n_2+1)(n_1+3)}$$

En otro orden de ideas, Sokal y Rohlf, exponen que la comparación de la eficiencia relativa no tiene mucho significado en sí misma, sino se consideran los costos relativos de los dos diseños. Claramente, si un diseño es dos veces más eficiente que otro (esto es, posee la mitad de la varianza que el primero), pero al mismo tiempo es diez veces más caro de realizar, no realizaríamos el más caro. Entonces, resulta obvio que para analizar correctamente la determinación de la eficiencia relativa se deben considerar las funciones de costo total de los dos diseños implicados.

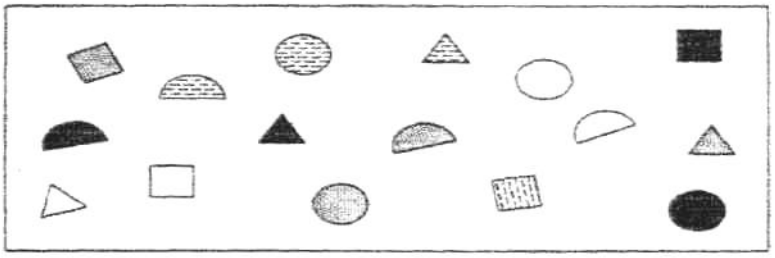
Capítulo IV

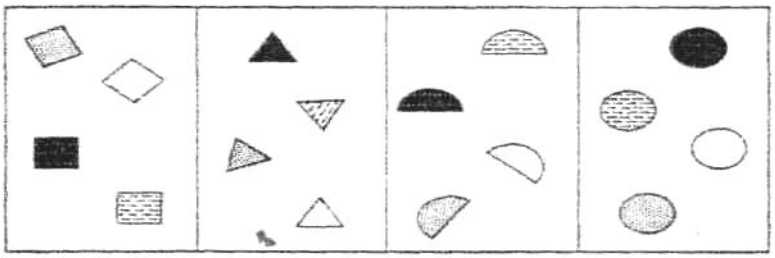
Diseños de cuadrados latinos y diseños afines

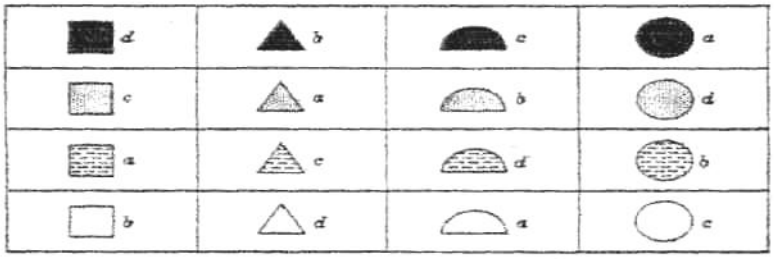
Estos diseños clásicos son una extensión lógica y natural del diseño en bloques completos al azar y poseen una serie de características muy similares, por tanto, se explicarán en conjunto.

Estos planes experimentales tienen como propósito el control de la variación del material experimental. A medida que éste se hace más heterogéneo es preciso controlar la variación bloqueando por cada característica que varíe. Así, el diseño de cuadrado latino (o doble bloqueo) se trata, como su nombre lo indica, de controlar dos fuentes de variación existentes, reconocidas por el investigador, entre las unidades experimentales. El investigador al controlar estas dos fuentes logra reducir la varianza del error posibilitando la expresión de la diferencia entre los tratamientos, si es que existe.

El Esquema del Diseño es:



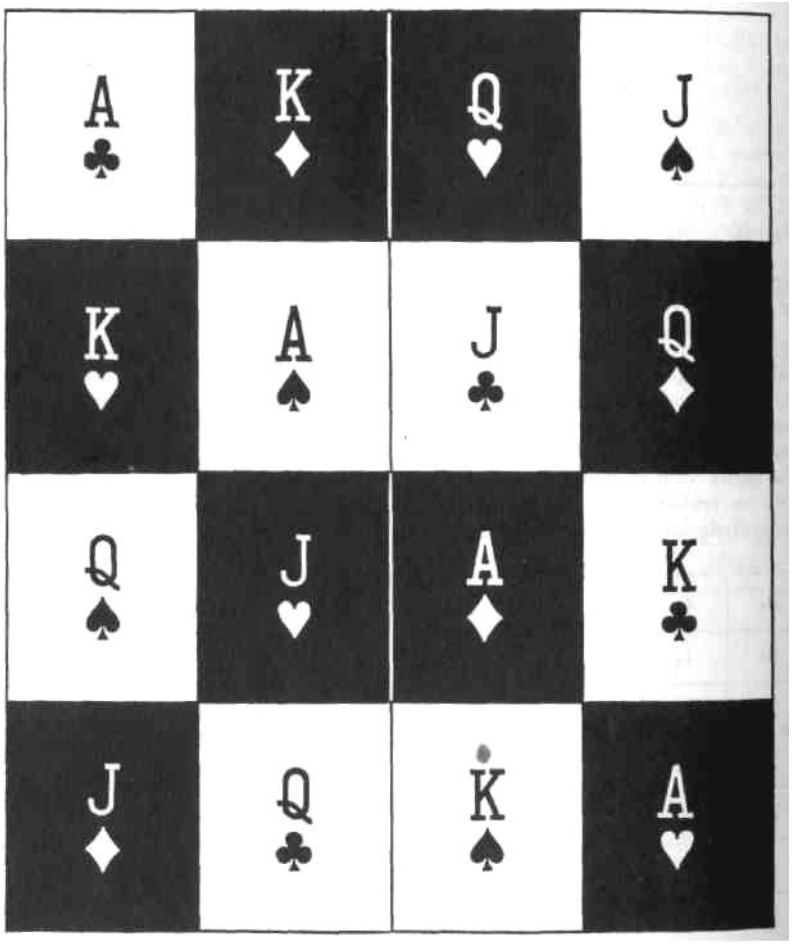




En el esquema anterior, en el recuadro superior, se aprecia un conjunto de unidades experimentales heterogéneas en dos sentidos (forma y color). Estas características son útiles al investigador para realizar el bloqueo. En el recuadro intermedio, el investigador ha procedido a bloquear según la forma de la figura geométrica. Y en el recuadro inferior el investigador ha completado el bloqueo, bloqueando según el color de la figuras geométricas, que es la otra característica resaltante y finalmente procede a asignar al azar los tratamientos a las unidades experimentales. Los tratamientos están representados por las letras latinas lo que da nombre al diseño.

Es importante destacar que cada tratamiento aparece sólo una vez por columna y por fila, permitiendo que cada tratamiento sea probado por igual según las dos fuentes de variación consideradas. Por otro lado, al seguir aumentando la variación surge otro plan experimental llamado Diseño de Cuadrado Greco-Latino, en este se desean controlar tres fuentes de variación y probar los tratamientos. En este cuadrado en realidad hay una superposición de dos cuadrados latinos.

El arreglo es como sigue:



El esquema sugiere una fuente de variación controlada por los iconos de las cartas o barajas (corazón blanco y negro, trébol blanco y negro, rombo blanco y negro y hoja negro y blanco), la otra fuente es el color (blanco y negro), las filas controlan otra fuente de variación, así como, las columnas. Esto produce un control de tres fuentes (filas, columnas, iconos de las cartas). Los tratamientos están representados por las letras latinas. Se aprecia que la permutación de filas y columnas dará origen a otro cuadrado para probar cuatro tratamientos.

Los diseños de cuadrado latinos en general presentan la dificultad de poseer muy pocos grados de libertad para estimar el error experimental. Esto ha obligado a los investigadores a utilizar como estrategia la combinación de una serie de cuadrados latinos del mismo orden en un mismo experimento. Cada aplicación (cada

cuadrado) constituye una réplica, esta estrategia es útil para aumentar los grados de libertad del error a este diseño se le conoce como: Diseño de Cuadrado Latino replicado.

Sea por ejemplo, una experimento donde se utilizó una batería de 4 cuadros latinos de orden 3x3:

Primer cuadrado

A	B	C
C	A	B
B	C	A

Segundo cuadrado

A	B	C
B	C	A
C	A	B

Tercer cuadrado

A	B	C
C	A	B
B	C	A

Cuarto cuadrado

A	B	C
B	C	A
C	A	B

Para este arreglo la tabla de ANOVA para un diseño de cuadrado latino replicado es:

Fuentes de variación	Grados de libertad	
Cuadrados	4-1= 3	
Filas (cuadrados)	4(3-1)= 8	
Columnas (cuadrados)	4(3-1)= 8	27
Tratamientos	3-1= 2	
Interacción tratamientos x cuadrados	(3-1)(4-1)= 6	
Error experimental	(36-1)-27= 8	
Total	(36-1)= 35	

Modelo Aditivo Lineal

El modelo aditivo lineal para este experimento es:

$$Y_{ijklm} = \mu + S_i + F(S)_{j/i} + C(S)_{k/i} + \tau_l + (\tau xS)_{i.l} + \varepsilon_{m/ijkl}$$

Donde:

Y_{ijklm} : es la respuesta

μ : media general

S_i : cuadrados latinos

$F(S)$: filas dentro de cuadrados

C(S): columnas dentro de cuadrados

τ : efecto del tratamiento

τxS : interacción del tratamiento y los cuadrados

ε : error experimental

Como resulta obvio para un diseño de cuadrado latino (con un solo cuadrado):

Fuentes de Variación	Grados de libertad
TRAT	3-1 =2
Filas	3-1=2
Columnas	3-1=2
Error Exp.	8-6=2
Total	(3x3)-1=8

Se aprecian dos grados de libertad para el error experimental en comparación con los 8 grados del diseño en cuadrado latino replicado. Sukhatme y Panse, establecen que los grados de libertad para estos diseño deben ser cuando menos 12 (doce) grados de libertad para el error, debido a que si inspeccionamos las tablas de F de Fisher-Snedecor observamos que para valores menores de 12 grados de libertad del error los valores de F tabulados presentan amplias variaciones en magnitud, mientras que para valores superiores se presentan valores de F tabulados estables.

Existe un plan de diseño experimental que incluye cuatro fuentes de variación se llama Cuadrado Hiper-Greco-Latino. Donde se aprecia una fuente controlada por las filas de cuadrado, otra por las columnas, otra por las letras latinas y griegas.

El esquema para este plan es:

	W1	W2	W3	W4
S1	$C\gamma 8$	$B\beta 6$	$A\alpha 5$	$D\delta 6$
S2	$A\delta 4$	$D\alpha 3$	$C\beta 7$	$B\gamma 3$
S3	$D\beta 5$	$A\gamma 6$	$B\delta 5$	$C\alpha 6$
S4	$B\alpha 6$	$C\delta 10$	$D\gamma 10$	$A\beta 8$

En este arreglo los tratamientos están representados por las letras latinas, los sub-tratamientos por las letras griegas, las filas controlan una fuente de variación y las columnas otra.

Ejemplo numérico de un diseño de cuadrado greco-latino

Sea una investigación para producir una mejora en un tipo de alimento para pollos bebés, se añaden cuatro cantidades diferentes de cada uno de dos químicos a los ingredientes básicos de la ración.
Las cantidades diferentes del primer ingrediente químico se indican como A,B,C y D, mientras que las del segundo por $\alpha, \beta, \gamma, \delta$. El

alimento de suministra a los pollos bebés ordenados en grupos, de acuerdo a 4 pesos iniciales diferentes (W1,W2,W3 y W4) y cuatro razas o especies diferentes (S1,S2,S3 y S4). Los incrementos de peso, por unidad de tiempo, se presentan en el cuadro dado anteriormente. Se pide: (a) Realizar el ANOVA, (b) Presentar el modelo aditivo lineal, (c) Expresar conclusiones.

Base de datos

Filas	Col	LAT	GRI	PESO
1	1	3	3	8
1	2	2	2	6
1	3	1	1	5
1	4	4	4	6
2	1	1	4	4
2	2	4	1	3
2	3	3	2	7
2	4	2	3	3
3	1	4	2	5
3	2	1	3	6
3	3	2	4	5
3	4	3	1	6
4	1	2	1	6
4	2	3	4	10
4	3	4	3	10
4	4	1	2	8

El modelo aditivo lineal es:

$$Y_{ijklm} = \mu + F_i + C_j + L_k + G_l + \varepsilon_{m/ijkl}$$

Donde:

Y_{ijklm} : variable respuesta

μ : media general

F_i : efecto de fila

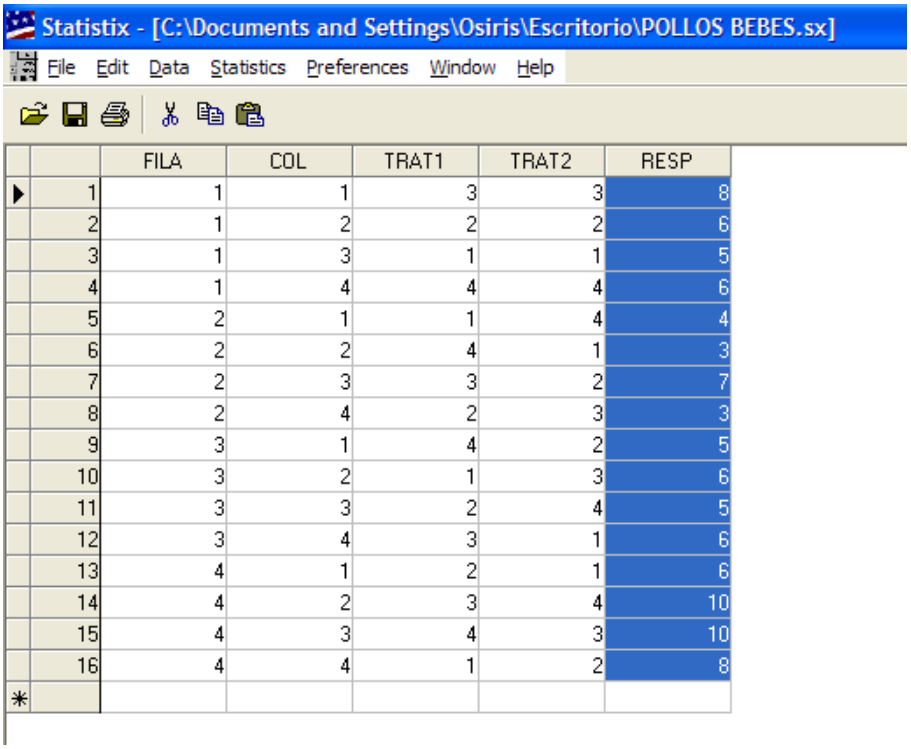
C_j : efecto de columna

L_k : efecto del tratamiento

G_l : efecto del subtratamiento

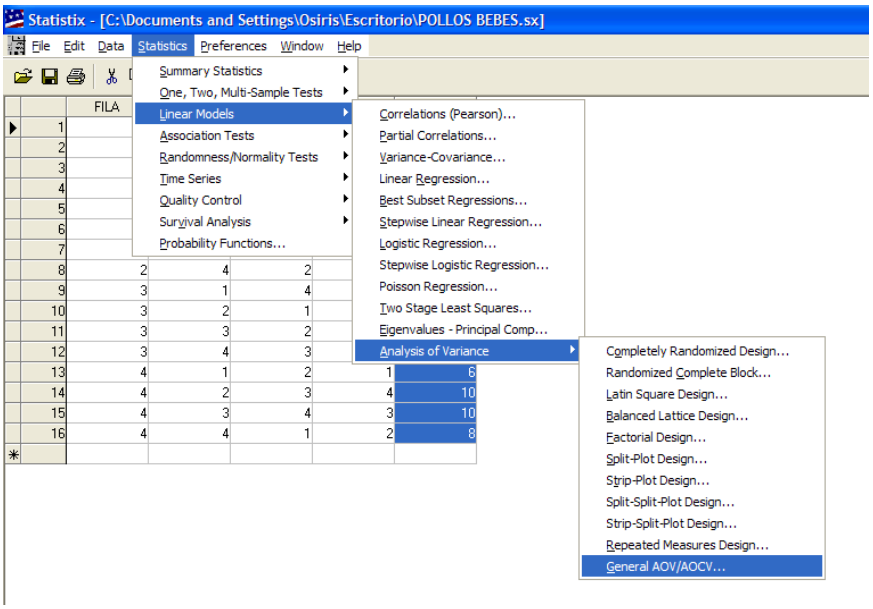
$\varepsilon_{m/ijkl}$: efecto del error experimental

Matriz de datos en el Statistix

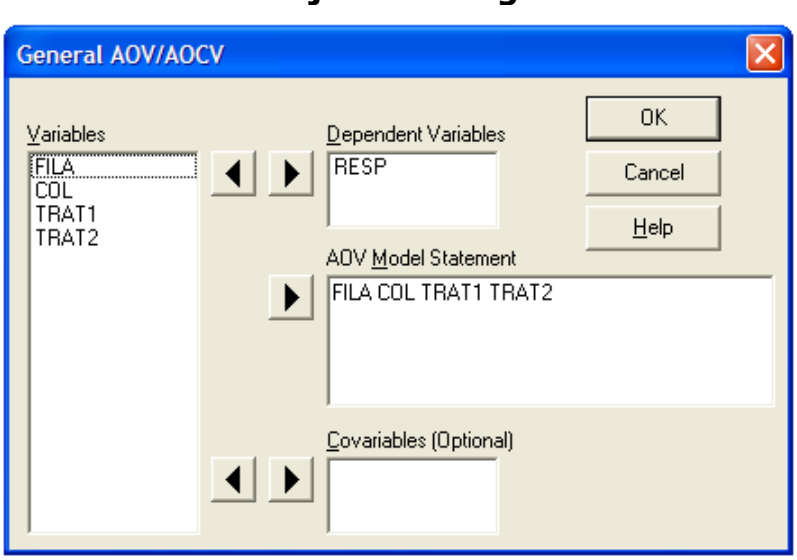


	FILA	COL	TRAT1	TRAT2	RESP
1	1	1	3	3	8
2	1	2	2	2	6
3	1	3	1	1	5
4	1	4	4	4	6
5	2	1	1	4	4
6	2	2	4	1	3
7	2	3	3	2	7
8	2	4	2	3	3
9	3	1	4	2	5
10	3	2	1	3	6
11	3	3	2	4	5
12	3	4	3	1	6
13	4	1	2	1	6
14	4	2	3	4	10
15	4	3	4	3	10
16	4	4	1	2	8

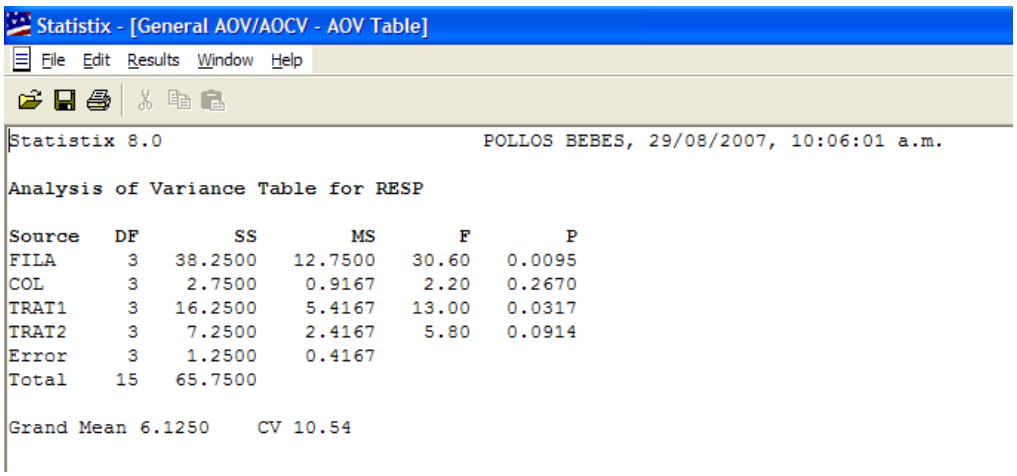
Secuencia de opciones en el Statistix



Caja de dialogo



Cuadro de Anova



Statistix - [General AOV/AOCV - AOV Table]

File Edit Results Window Help

Statistix 8.0 POLLOS BEBES, 29/08/2007, 10:06:01 a.m.

Analysis of Variance Table for RESP

Source	DF	SS	MS	F	P
FILA	3	38.2500	12.7500	30.60	0.0095
COL	3	2.7500	0.9167	2.20	0.2670
TRAT1	3	16.2500	5.4167	13.00	0.0317
TRAT2	3	7.2500	2.4167	5.80	0.0914
Error	3	1.2500	0.4167		
Total	15	65.7500			

Grand Mean 6.1250 CV 10.54

Conclusiones:

(a) Existe diferencia significativa en el incremento de peso por razas.

(b) Existe diferencia significativa para el tratamiento 1, es decir, el primer ingrediente.

(c) El incremento de peso no varió (no fue significativo) por peso inicial del pollito ni por el segundo ingrediente.

Capítulo V

Experimentos factoriales

Se llaman Experimentos Factoriales a aquellos experimentos en los que se estudia simultáneamente dos o más factores, y donde los tratamientos se forman por la combinación de los diferentes niveles de cada uno de los factores.

Los experimentos factoriales en si no constituyen un diseño experimental si no un Diseño de Tratamiento (un arreglo de tratamiento es una disposición geométrica de ellos bien en el espacio o en el tiempo y que deben ser llevados en cualquiera de los diseños experimentales clásicos tal como el Diseño Completo al Azar, el Diseño en Bloques Completos al Azar, el Diseño en Cuadrado Latino.

Los experimentos factoriales se emplean en todos los campos de la investigación, son muy útiles en investigaciones exploratorias en las que poco se sabe acerca de muchos factores. Muy frecuentemente usados en investigaciones comparativas.

Ventajas:

1.- Permiten estudiar los efectos principales, efectos de interacción de factores, efectos simples y efectos cruzados y anidados.

2.- Todas las unidades experimentales intervienen en la determinación de los efectos principales y de los efectos de interacción de los factores, por lo que el número de repeticiones es elevado para estos casos.

3.- El número de grados de libertad para el error experimental es alto, comparándolo con los grados de libertad de los experimentos simples de los mismos factores, lo que contribuye a disminuir la varianza del error experimental, aumentando por este motivo la precisión del experimento.

Desventajas:

1.- Se requiere un mayor número de unidades experimentales que los experimentos simples y por lo tanto se tendrá un mayor costo y trabajo en la ejecución del experimento.

2.- Como en los experimentos factoriales cada uno de los niveles de un factor se combinan con los niveles de los otros factores; a fin de que exista un balance en el análisis estadístico se tendrá que *algunas de las combinaciones no tiene interés práctico* pero deben incluirse para mantener el balance.

3.- El análisis estadístico es más complicado que en los experimentos simples y la interpretación de los resultados se hace más difícil a medida de que aumenta el número de factores y niveles por factor en el experimento.

Conceptos generales:

Factor.- Es un conjunto de tratamientos de una misma clase o característica. Ejemplo: tipos de riego, dosis de fertilización, variedades de cultivo, manejo de crianzas, métodos de enseñanza, tipos de liderazgo, tipos raciales, etc.

Factorial.- Es una combinación de factores para formar tratamientos.

Niveles de un factor.- Son los diferentes tratamientos que pertenecen a un determinado factor. Se acostumbra simbolizar algún elemento "i" por la letra minúscula que representa al factor y el valor del respectivo subíndice.

Ejemplo:

A: Tipos de riego: Secano, Goteo, Aspersión

Niveles: a_0 a_1 a_2

Tipos de factores:

1.- Factores Cuantitativos.

2.- Factores Cualitativos.

1.- Factores cuantitativos.- Son aquellos factores cuyos niveles son cantidades numéricas.

Ejemplo:

Factor A : Dosis de fertilización

Niveles : 10 Kg/Ha (a_o), 20Kg/Ha (a_1), 30Kg/Ha (a_2).

2.- Factores cualitativos.- Son aquellos factores cuyos niveles son procedimientos, o cualidades o atributos.

Ejemplo:

Factor A: Variedades de cultivo

Niveles : Variedad 1, Variedad 2.

Interacción

Es el efecto combinado de dos o más factores.

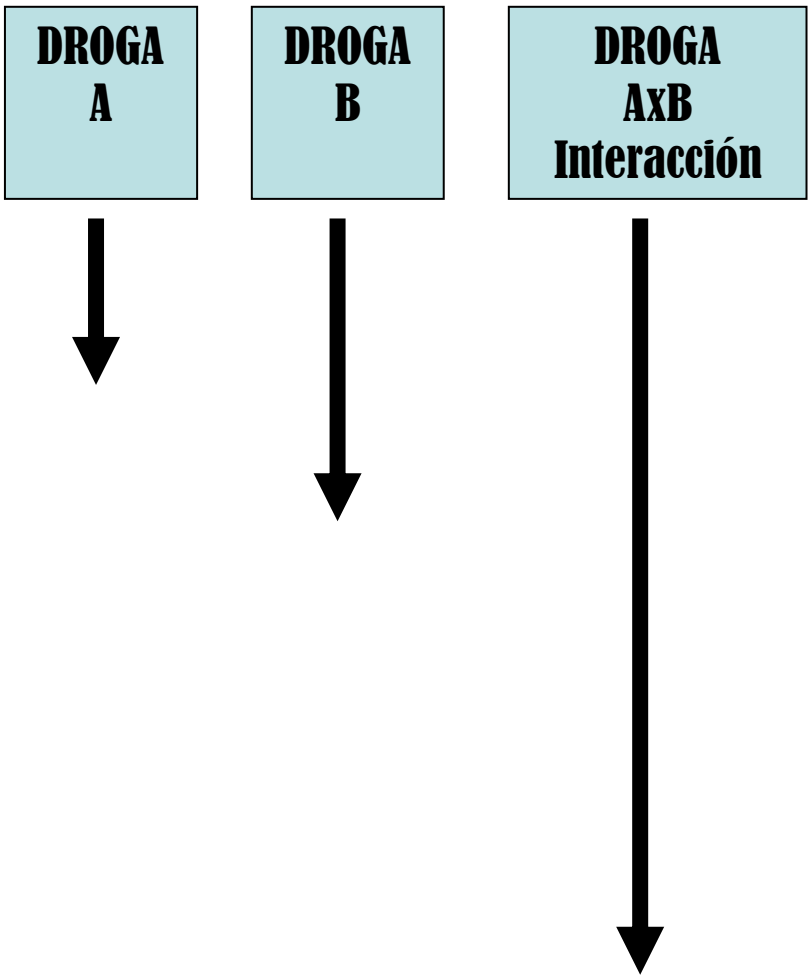
Es la combinación de dos o más variables independientes para generar un efecto diferente al que ellos tienen cuando actúan independientemente.

El experimento factorial se planifica con la intención exprofesa de medir la interacción y evaluarla. La interacción puede ser de tres tipos: Sinergismo, Antagonismo, Aditivo. Es análoga a la acción de una droga.

Interacción por sinergismo

Las dos drogas (factores) se combinan y generan un efecto muy superior al que ellas exhiben por separado:

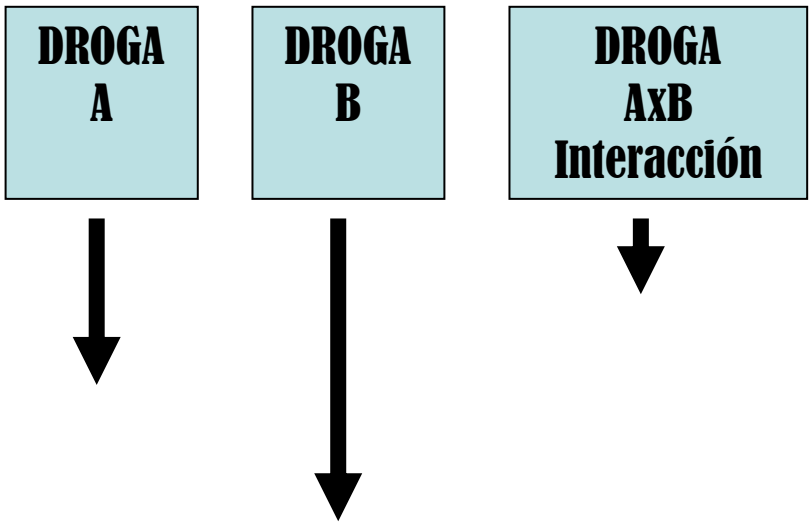
Interacción por Sinergismo



Interacción por antagonismo

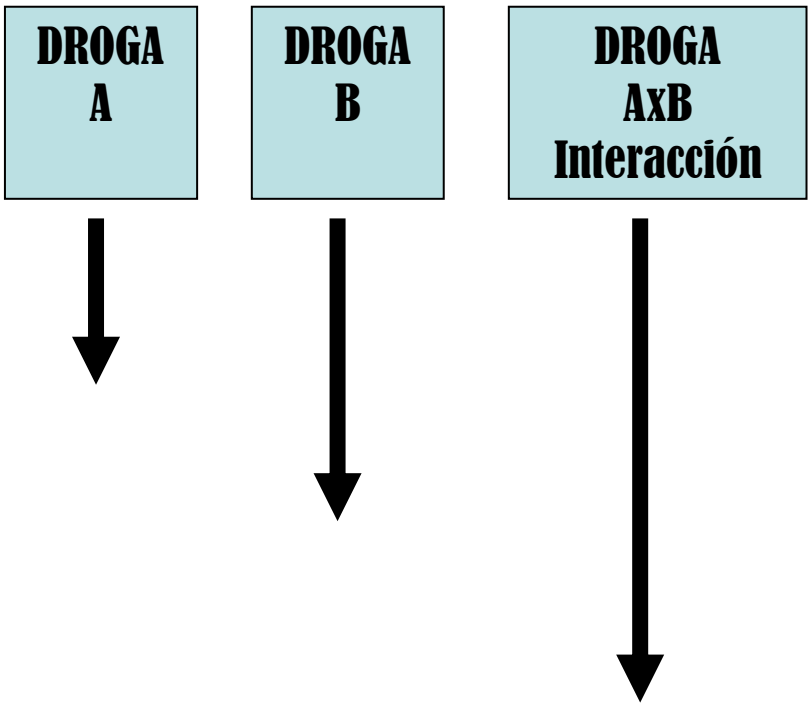
Las drogas (o factores) se combinan para producir un efecto inferior al que ellas exhiben por separado:

Interacción por Antagonismo



Interacción Aditiva: Las drogas (o factores) se combinan y producen un efecto igual a la suma de los efectos independientes de cada factor:

Interacción Aditiva



Algunos autores se refieren a la interacción como el efecto que se gana o se pierde cuando se combinan dos o más factores y hablan en

dichos caos de interacción positiva (cuando se gana) e interacción negativa (cuando se pierde). En economía, se habla de incremento y decremento. O incremento positivo y negativo.

Ejemplo de formación de factoriales:

Sea los factores A y B con sus respectivos niveles:

Factor A: a_0 a_1 a_2

Factor B: b_0 b_1

La combinación de los niveles de los factores será:

a_0 a_1 a_2

b_0 b_1 b_0 b_1 b_0 b_1

$\{a_0 b_0 \quad a_0 b_1 \quad a_1 b_0 \quad a_1 b_1 \quad a_2 b_0 \quad a_2 b_1\}$ tratamientos

Al combinar ambos factores (A y B) se tiene:

3 x 2 = 6 tratamientos para ser evaluados

niveles de A x niveles de B

Si cada tratamiento se aplica a 4 unidades experimentales, se requiere 24 unidades experimentales, para realizar el experimento:

Repeticiones	a_0 b_0	a_0 b_1	a_1 b_0	a_1 b_1	$a_1 b_0$	a_2 b_1
1						
2						
3						
4						

Formación de factoriales:

En la información de factoriales, se debe tener presente lo siguiente:

1.- Que factores deben incluirse.

2.- Que factores son fijos (modelo I) y que factores son al azar (modelo II).

3.- Cuántos niveles se tiene por factor.

4.- Si son factores cuantitativos, cual debe ser el espaciamiento entre los niveles del factor.

Por ejemplo:

0%, 5% y 10% de nitrógeno, significa igual espaciamiento.

Tipos de experimentos factoriales:

Los experimentos factoriales para un determinado diseño se diferencian entre si, por el número de factores y por la cantidad de niveles de estos factores que intervienen en el experimento.

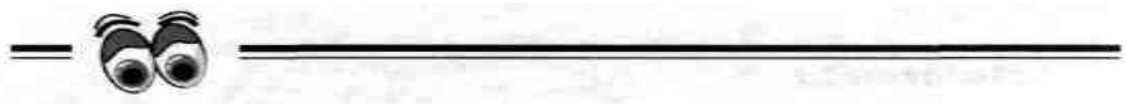
Para simbolizar se usa la letra del factor:

pA x qB dos factores "A y "B", con "p" niveles para "A" y "q" niveles para "B"

2A 2B = 2A x 2B ó 2 x 2 ó 2^2 Número de niveles de cada factor lo cual permite evaluar (4 tratamientos).

3A 3B = 3A x 3B ó 3 x 3 ó 3^2 Número de niveles (9 tratamientos).

4A 4B = 4A x 4B ó 4 x 4 ´ó 4^2 Número de niveles (l6 tratamientos).



Efectos de los experimentos factoriales:

1.- Efecto principal.- Es una medida del cambio en el promedio entre los niveles de un factor, promediado sobre los diferentes niveles del otro factor. Ejemplo: Dosis de Nitrogeno en las U.E.

2.- Efecto interacción.- Es una medida de cambio que expresa el efecto adicional resultante de la influencia combinada de dos o más factores.

Ejemplo: Efecto conjunto de nitrógeno y fósforo.

3.- Efecto simple.- Es una medida de cambio en los promedios de los niveles de un factor, manteniendo constante, uno de los niveles del otro factor.

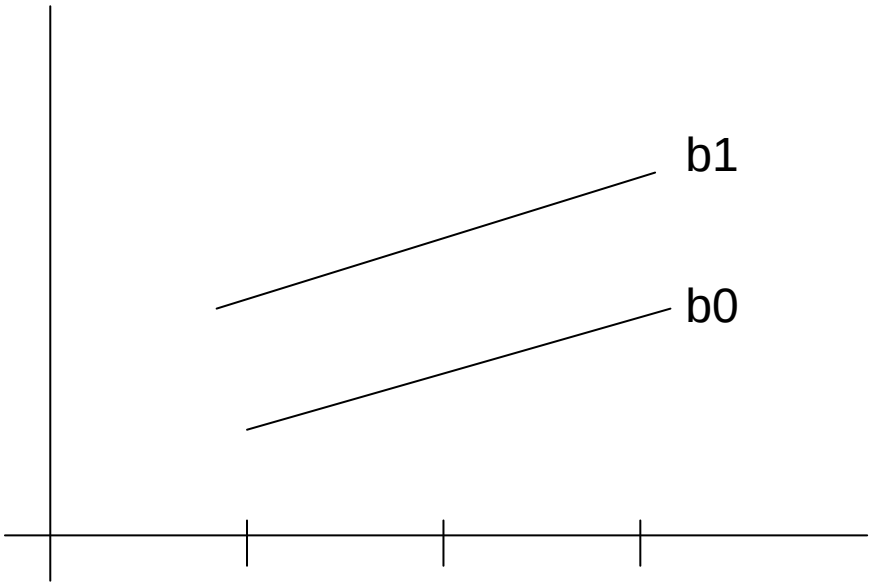
Ejemplo: Efecto de nitrógeno ante la presencia de 5% de fósforo.

Gráfico de la interacción:

La interacción de los factores se representa gráficamente; la tendencia indica el grado de interacción entre los factores, la cual aumenta a medida que las líneas tiendan a cruzarse.

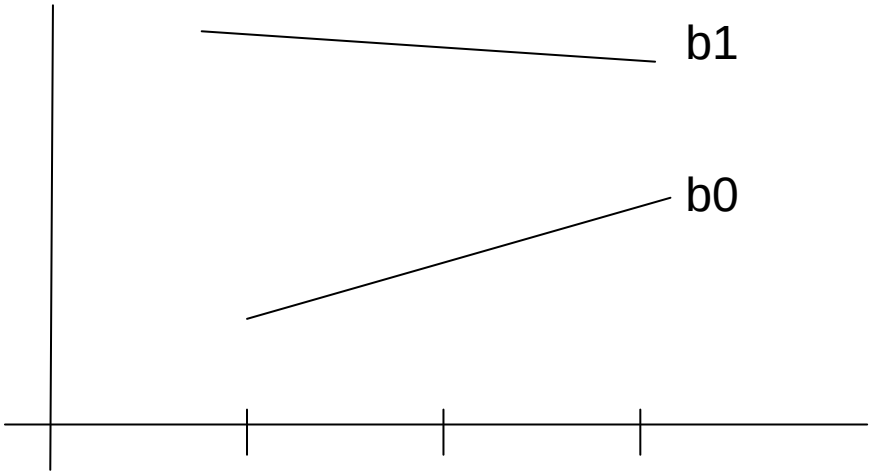
En los siguientes gráficos se muestran los casos posibles de interacción en dos factores: A con 3 niveles y B con 2 niveles. En el eje "X" se registra los niveles de A y en el eje "Y" los promedios de la interacción de "A" y "B". Los puntos son unidos por una línea de tendencia, para cada nivel de "B".

Sin interacción:



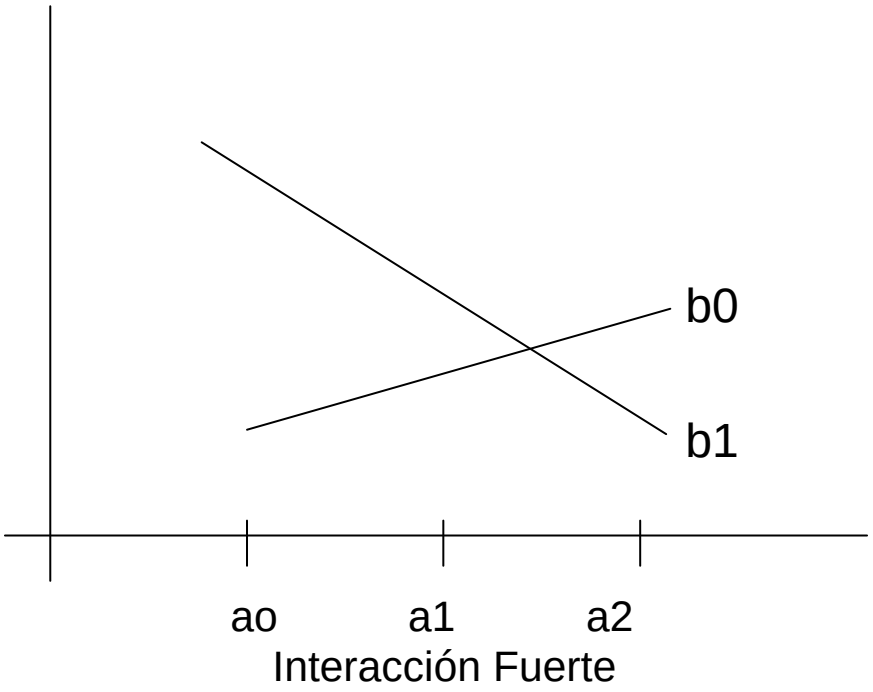
Sin interacción
Líneas Paralelas

Interacción débil:



Interacción Débil
Tendencia a entrecruzarse las líneas P

Interacción Fuerte:



Interacción Fuerte

Las líneas de Tendencia se entrecruzan

Experimento factorial (pxq) conducido en DBCA

Ejemplo práctico:

Con el objeto de averiguar la estabilidad de la vitamina C en concentrado de jugo de naranja (congelado, reconstituido) que se almacena en un refrigerador por un periodo de hasta una semana, se probaron 3 marcas del jugo de naranja, a 3 diferentes tiempos. Estos últimos se refieren al número de días que transcurren desde que el jugo de naranja se mezcla hasta que se somete a la prueba. La información se recogió de 4 diferentes muestras, provenientes de 4 diferentes operarios. Los resultados expresados en miligramos de ácido ascórbico por litro se presentan a continuación:

Factor M : (marca): M_1, M_2, M_3

Factor T : (tiempo, días) : T_1(3 días), T_2(5 días), T_3(7 días)

Bloques : (Operarios) : B_1, B_2, B_3, B_4

	M1			M2			M3			
BLOQ	T1	T2	T3	T1	T2	T3	T1	T2	T3	TOTAL
1	52.6	56.0	52.5	49.4	48.8	48.0	42.7	49.2	48.5	447.7
2	54.2	48.0	52.0	49.2	44.0	47.0	48.8	44.0	43.4	430.6
3	49.8	49.6	51.8	42.8	44.0	48.2	40.4	42.0	45.2	413.8
4	46.5	48.4	53.6	53.2	42.4	49.6	47.6	43.2	47.6	432.1
TOTAL	203.1	202.0	209.9	194.6	179.2	192.8	179.5	178.4	184.7	1724.2
	M1 = 615.0			M2 = 566.6			M3 = 542.6			
	T1 = 577.2			T2 = 559.6			T3 = 587.4			

Donde:

i: 1, 2, 3 (Marca) j : 1, 2, 3 (Tiempo) k : 1, 2, 3, 4 (Operario)

$$\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{4} Y_{ijk}^2 = 83{,}110.68 \qquad \sum_{i=1}^{3}\sum_{j=1}^{3} Y_{ij.}^2 = 331{,}426.16 \qquad \sum_{i=1}^{3} Y_{i..}^2 = 993{,}675.32$$

$$\sum_{j=1}^{3} Y_{.j.}^2 = 991{,}350.76 \qquad \sum_{k=1}^{4} Y_{..k}^2 = 743{,}792.5 \qquad \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{4} Y_{ijk} = 1{,}724.2$$

Se pide:

a. Presente el Modelo Aditivo Lineal (M.A.L.) e interprete cada uno de sus componentes en términos del problema.
b. Realice el Análisis gráfico de la interacción. ¿Qué comentario puede Ud., realizar al respecto ?.
c. Realice el Análisis de Variancia. De sus conclusiones con alpha = 0.05
d. Realice la Prueba de Duncan para el Factor Marca de Jugo de Naranja. Use alpha = 0.05

Algunos diseños experimentales clásicos

Un diseño experimental es una regla que determina la asignación de las unidades experimentales a los tratamientos. Aunque los experimentos difieren unos de otros en muchos aspectos, existen

diseños estándar que se utilizan con mucha frecuencia. Algunos de los más utilizados son los siguientes:

1. Diseño completamente aleatorizado:

El experimentador asigna las unidades experimentales a los tratamientos al azar. La única restricción es el número de observaciones que se toman en cada tratamiento. De hecho si n_i es el número de observaciones en el i-ésimo tratamiento, $i = 1,...,I,$ entonces, los valores $n_1,n_2,...,n_I$ determinan por completo las propiedades estadísticas del diseño. Naturalmente, este tipo de diseño se utiliza en experimentos que no incluyen factores bloque y las unidades experimentales sean muy homogéneas.

El modelo matemático de este diseño tiene la forma:

$$\text{Respuesta} = \text{Constante} + \text{Efecto tratamiento} + \text{Error}$$

2. Diseño en bloques o con un factor bloque:

En este diseño el experimentador agrupa las unidades experimentales en bloques, a continuación determina la distribución de los tratamientos en cada bloque y, por último, asigna al azar las unidades experimentales a los tratamientos dentro de cada bloque.

En el análisis estadístico de un diseño en bloques, éstos se tratan como los niveles de un único factor de bloqueo, aunque en realidad puedan venir definidos por la combinación de niveles de más de un factor nuisance (factor de ruido)

El modelo matemático de este diseño es:

$$\text{Respuesta} = \text{Constante} + \text{Efecto bloque} + \text{Efecto tratamiento} + \text{Error}$$

El diseño en bloques más simple es el denominado **diseño en bloques completos,** en el que cada tratamiento se observa el mismo número de veces en cada bloque.

El diseño en bloques completos con una única observación por cada tratamiento se denomina **diseño en bloques completamente aleatorizado** o, simplemente, **diseño en bloques aleatorizado**.

Cuando el tamaño del bloque es inferior al número de tratamientos no es posible observar la totalidad de tratamientos en cada bloque y se habla entonces de **diseño en bloques incompletos**.

3. Diseños con dos o más factores bloque:

En ocasiones hay dos (o más) fuentes de variación lo suficientemente importantes como para ser designadas factores de bloqueo. En tal caso, ambos factores bloque pueden ser **cruzados o anidados.**

Los factores bloque están **cruzados** cuando existen unidades experimentales en todas las combinaciones posibles de los niveles de los factores bloques.

Diseño con factores bloque cruzados. También denominado **diseño fila-columna***,* se caracteriza porque existen unidades experimentales en todas las *celdas* (intersecciones de fila y columna).

El modelo matemático de este diseño es:

$$\text{Respuesta} = \text{Constante} + \text{Efecto bloque fila} + \text{Efecto bloque columna} +$$

$$\text{Efecto tratamiento} + \text{Error}$$

Los factores bloque están **anidados** si cada nivel particular de uno de los factores bloque ocurre en un único nivel del otro factor bloque.

4. Diseños con dos o más factores:

En algunas ocasiones se está interesado en estudiar la influencia de dos (o más) factores tratamiento, para ello se hace un diseño de filas por columnas. En este modelo es importante estudiar la posible interacción entre los dos factores. Si en cada casilla se tiene una única observación no es posible estudiar la interacción entre los dos factores, para hacerlo hay que replicar el modelo, esto es, obtener *k* observaciones en cada casilla, donde *k* es el número de réplicas.

El modelo matemático de este diseño es:

Para dos factores:

$$Y = A + B + A*B + \varepsilon$$

Para tres factores:

$$Y = A + B + C + A*B + A*C + B*C + A*B*C$$

Para cuatro factores:

$$Y = A + B + C + D + A*B + A*C + A*D + B*C + B*D + C*D + A*B*C + A*B*D + B*C*D + A*C*D + A*B*C*D$$

Generalizar los diseños completos a más de dos factores es relativamente sencillo desde un punto de vista matemático, pero en su aspecto práctico tiene el inconveniente de que al aumentar el número de factores aumenta muy rápidamente el número de observaciones necesario para estimar el modelo. En la práctica es muy raro utilizar diseños completos con más de 5 factores.

Un camino alternativo es utilizar **fracciones factoriales** que son diseños en los que se supone que muchas de las interacciones son nulas, esto permite estudiar el efecto de un número elevado de

factores con un número relativamente pequeño de pruebas. Por ejemplo, el diseño en **cuadrado latino,** en el que se supone que todas las interacciones son nulas, permite estudiar tres factores de k niveles con solo k^2 observaciones. Si se utilizase el diseño equilibrado completo se necesitan k^3 observaciones.

5. Diseños factoriales a dos niveles:

En el estudio sobre la mejora de procesos industriales (control de calidad) es usual trabajar en problemas en los que hay muchos factores que pueden influir en la variable de interés. La utilización de experimentos completos en estos problemas tiene el gran inconveniente de necesitar un número elevado de observaciones, además puede ser una estrategia ineficaz porque, por lo general, muchos de los factores en estudio no son influyentes y mucha información recogida no es relevante. En este caso una estrategia mejor es utilizar una técnica secuencial donde se comienza por trabajar con unos pocos factores y según los resultados que se obtienen se eligen los factores a estudiar en la segunda etapa.

Los **diseños factoriales** 2^k son diseños en los que se trabaja con k factores, todos ellos con dos niveles (se suelen denotar + y -). Estos diseños son adecuados para tratar el tipo de problemas descritos porque permiten trabajar con un número elevado de factores y son válidos para estrategias secuenciales.

Si k es grande, el número de observaciones que necesita un diseño factorial 2^k es muy grande ($n = 2^k$). Por este motivo, las **fracciones factoriales** 2^{k-p} son muy utilizadas, éstas son diseños con k factores a dos niveles, que mantienen la propiedad de ortogonalidad de los factores y donde se suponen nulas las interacciones de orden alto (se confunden con los efectos simples) por lo que para su estudio solo se necesitan 2^{k-p} observaciones

(cuanto mayor sea *p* menor número de observaciones se necesita pero mayor confusión de efectos se supone).

En los últimos años Taguchi ha propuesto la utilización de fracciones factoriales con factores a tres niveles en problemas de control de calidad industrial.

Diseños Anidados o Jerarquizados

Los diseños anidados o jerarquizados son arreglos factoriales de tratamientos, donde un factor se encuentra encajado (anidado o jerarquizado) dentro de otro y este a su vez dentro de otro. Esta disposición permite que se prueben ciertos niveles y no otros. Permite que se prueben sólo los niveles que incluye el investigador.

Si se concibe de esta forma se considera como una estrategia de control de la variación con la intención de reducir la variación debida al error experimental.

Suele confundirse con el arreglo factorial cruzado. Ejemplo sea un factorial de tres factores cruzados del tipo (2x2x3):

Factor a SEXO	Factor b VITAMINA	Factor c MINERAL	Variable respuesta Pesos= kilogramos
a1 (varón)	b1 (500)	c1 (20%) c2 (40%) c3 (80%)	2,3,4 1,2,3 5,2,3
	b2 (1000)	c1 (20%) c2 (40%) c3 (80%)	5,2,3 1,4,6 2,5,6
a2 (mujer)	b1 (500)	c1 (20%) c2 (40%) c3 (80%)	3,3,3 5,4,3 2,3,1
	b2 (1000)	c1 (20%) c2 (40%) c3 (80%)	8,9,5 2,3,4 4,5,6

En un ejemplo de nutrición puede considerarse: factor a (ao=varón, a1=mujer), factor b (b1=vitamina 500 UI, b2= vitamina 1000 UI) y un tercer factor c (c1=mineral 20%, c2=mineral 40% y c3=mineral al 80%).

La variable respuesta es el aumento de peso experimentado por el sujeto en un período de 4 meses.

De esta forma se dice que b está cruzado con a por ser los mismos niveles; c está cruzado con b por ser los mismos niveles; c está cruzado con a por ser los mismos niveles.

Pero si introducimos un cambio en el diseño anterior así, cambiamos o queremos probar intencionalmente niveles de vitaminas y minerales diferentes tendremos:

Factor a (SEXO)	Factor b (VITAMINA)	Factor c (MINERAL)	Variable respuesta Peso =Kgr
a1 (varón)	b1 (500)	c1 (20%) c2 (40%) c3 (80%)	2,3,4 1,2,3 5,2,3
	b2 (1000)	c1 (25%) c2 (45%) c3 (65%)	5,2,3 1,4,6 2,5,6
a2 (mujer)	b1 (300)	c1 (12%) c2 (13%) c3 (14%)	3,3,3 5,4,3 2,3,1
	b2 (2000)	c1 (30%) c2 (40%) c3 (50%)	8,9,5 2,3,4 4,5,6

En este nuevo arreglo los factores vitaminas y minerales que están encajados por sexo no son los mismos. Esta disposición nos lleva a afirmar que el diseño está anidado: b dentro de a y c dentro de b y a. Las dos disposiciones experimentales conllevan a análisis estadísticos diferentes porque el primero incluye interacciones y el segundo no.

El modelo aditivo lineal del primer ensayo es:

$$Y = \mu + A + B + C + (AxB) + (AxC) + (BxC) + (AxBxC) + \varepsilon$$

Donde:

Y= la variable respuesta (el aumento de peso)

$\mu =$ media general de la respuesta

A=efecto del sexo

B=efecto de la vitamina

C=efecto del mineral

AxB=efecto de la interacción sexo por vitamina

AxC= efecto de la interacción sexo por mineral

BxC= efecto de la interacción vitamina por mineral

AxBxC= efecto de la interacción sexo por vitamina por mineral

$\varepsilon =$ error experimental

Para el segundo arreglo, el modelo aditivo lineal es:

$$Y = \mu + A + B(A) + C(AxB) + \varepsilon$$

Donde:

Y= la variable respuesta (el aumento de peso)

$\mu =$ media general de la respuesta

A=efecto del sexo

B(A)=efecto de la vitamina dentro de sexo

C(AxB)=efecto del mineral dentro de sexo y mineral

$\varepsilon =$ error experimental

Realizar el tratamiento estadístico de los dos ensayos y verificar las diferencias que existen entre ambos ANOVAS.

El Diseño es: Experimental Riguroso o Genuino.

Características y Propiedades:

(a) Existe Control de todos los factores de validez interna

(b) Existe aleatorización

(c) Hay replicación

Ejemplo de aplicación de un factorial

Un fabricante de alimentos prueba varias composiciones de un jugo de naranja. Se prueban tres niveles de dulzura (alto, medio, bajo) en combinación con dos niveles de acidez (bajo, alto) y dos coloraciones Natura (1) y Artificial (0). Seis personas califican cada combinación en una escala del 1 al 99. Las calificaciones son las siguientes:

Dulzura	Acidez	Color	Calificación						
1	1	1	40		35	45	42	48	45
1	1	0	62		56	60	54	60	65
1	2	1	38		32	50	40	46	35
1	2	0	60		50	48	61	55	53
2	1	1	45		56	50	45	55	48
2	1	0	72		56	63	75	67	68
2	2	1	47		53	46	52	56	47
2	2	0	60		66	56	64	72	70
3	1	1	35		50	35	40	43	35
3	1	0	56		48	52	45	59	52
3	2	1	25		35	30	24	34	34
3	2	0	40		36	32	31	34	38

Matriz de Datos en Statistix

Cuando vamos a crear la matriz de datos debemos tener cuidado en crear una variable para cada factor y prever una variable adicional para las interacciones (combinaciones en este caso: 12 en total):
La matriz de datos aparece en el CD-ROM en el archivo: JUGO.SX.
La secuencia de instrucciones es:

STATISTICS>LINEAR MODELS>ANALISIS DE VARIANZA>GENERAL AOV/AOCV

Caja de dialogo

Incluye la variable dependiente (aquí se declara (la calificación del jugo) y en las instrucciones del modelo de ANOVA incluiremos los factores principales y sus interacciones, es decir:

Variable dependiente: califica

Instrucciones del Modelo: A B C A*B A*C B*C A*B*C

O TAMBIÉN: ALL(A B C) instrucción abreviada que genera todas y cada una de las interacciones, así como los efectos principales.

El resumen de los resultados de las pruebas F es:

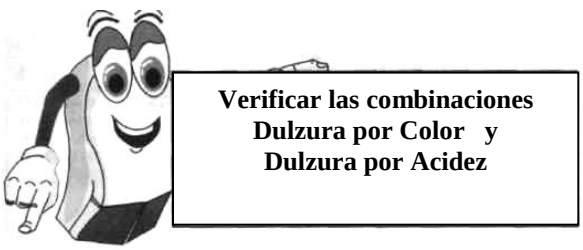
Fuentes de variación	gl	Sumas de cuadrados	Valor F	Valor P
Dulzura	2	4149,53	75,51	0,0001
Acidez	1	624,22	22,72	0,0001
Dulzura*Acidez	2	488,53	8,89	0,0004
Color	1	3200,00	116,46	0,0001
Dulzura*Color	2	203,08	3,70	0,0307
Acidez*Color	1	80,22	2,92	0,0927
Dulzura*Acidez*Color	2	24,19	0,44	0,6459

Conclusiones

(1) Las calificaciones del jugo variaron por: Dulzura, Acidez, Color y por la interacción Dulzura*Color y Dulzura*Color.

(2) Las calificaciones del jugo fueron iguales por: Acidez*Color y Dulzura, Acidez y Color.

(3) Es conveniente verificar a continuación ¿Cuál combinación es mejor, entre Dulzura y Color? ¿Cuál combinación es mejor, entre Dulzura y Acidez?



Capítulo VI

Análisis de Covarianza

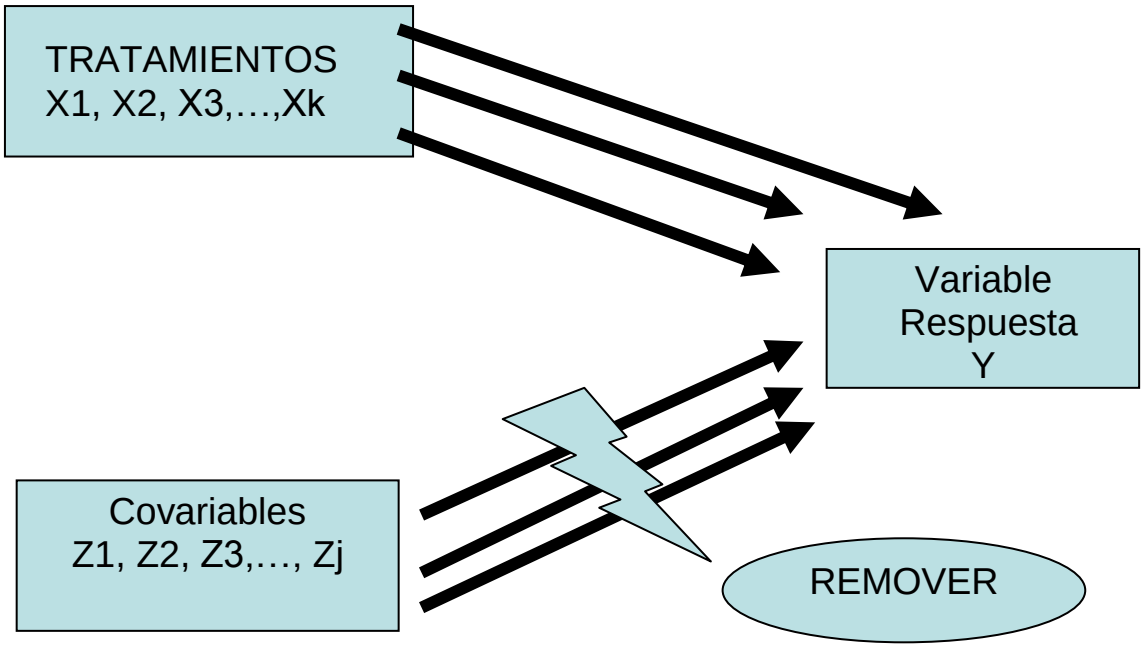
El análisis análisis de covarianza es un procedimiento muy importante en experimentación, pero lamentablemente no se usa con frecuencia. Utiliza el análisis de varianza y el de regresión para eliminar la variabilidad que existe en la variable independiente X (o covariable); también ajusta medias de tratamiento y así estima mucho mejor el efecto de la variable independiente (tratamiento) sobre la variable dependiente.

El análisis de Covarianza se define como una técnica estadística empleada cuando no se puede controlar una o más variable(s) extraña(s).

Filosofía del análisis

En el análisis de covarianza se contemplan básicamente tres juegos de variables. El primer juego de variables está constituido por las variables independientes (representando las condiciones experimentales o tratamientos que se quieren probar); Un segundo juego de variables independientes (que representan las variables sobre las cuales no se puede ejercer control, son variables extrañas) y estos dos juegos de variables actúan sobre la variable dependiente o variable respuesta provocando un efecto. El análisis de covarianza lo que hace es separar el efecto debido a los tratamientos de aquel debido a las variables extrañas, es decir, corregir la respuesta eliminando la influencia de las variables extrañas (llamadas también Covariables, por el hecho de variar conjuntamente con los tratamientos afectando la variable respuesta)

Esquemáticamente se puede representar así:



Después de remover el efecto de las covariables se puede verificar si hay efecto de los tratamientos y, de haber efecto, se le atribuye a los tratamientos o condiciones experimentales que se están probando.

Aplicaciones del análisis de covarianza (ANCOVA)

En el campo de ciencias de la salud:

- Para estudiar la relación entre edad (covariable) y la concentración de colesterol (variable dependiente) por estados (tratamientos)
- Para estudiar la relación entre edad (covariable) y peso (covariable) sobre el rendimiento (variable dependiente) para varias dietas (tratamientos)
- Para estudiar la relación entre peso inicial (covariable) y consumo(covariable) sobre el aumento de peso (variable dependiente) por razas (tratamientos)
- Para estudiar la relación entre calorias consumidas (covariable) y edad (covariable) sobre el incremento de peso (variable

dependiente) por país (tratamientos) en distintas épocas del año (bloque)

En el campo de la agronomía:

- Para estudiar la relación entre tamaño de la mazorca (covariable) y rendimiento por hectárea (variable dependiente) por variedad de maíz (tratamientos)
- Para estudiar la relación entre número de plantas (covariable) y rendimiento por parcela (variable dependiente) por variedad de sorgo (tratamientos)
- Para estudiar la relación entre fósforo inorgánico e inorgánico (covariables) y el rendimiento por parcela (variable dependiente) por tipo de suelo (tratamientos)

En el campo de la educación:

- Para estudiar la relación entre métodos de enseñanza, estilos cognitivos, edad y nivel de entrada (tratamientos) y personalidad, coeficiente intelectual y actitud científica (covariables) sobre el rendimiento en física y matemática (variables dependientes)

- Para estudiar la relación entre autoestima (variable dependiente) y el clima familiar (X=covariable) para diversos grupos familiares "con padre, con padrastro, sin padre" (tratamientos)

En el campo de la gerencia:

- Para estudiar la relación entre desempeño (variable dependiente) y el clima organizacional (X1=covariable) con tipo de liderazgo (X2=covariable) en diversos departamentos de la empresa (tratamientos)

Algunas consideraciones ligadas al ANCOVA

- Teóricamente no hay límites para el número de covariables que pueden usarse, pero la práctica ha demostrado que 4 ó 5 covariables son suficientes; más allá de este número se generán problemas de colinealidad.
- Las covariables son variables independientes que afectan la respuesta pero sobre las cuales no se ha podido ejercer control.
- La variable independiente X (covariable) es una observación hecha en cada unidad experimental antes de aplicar los tratamientos, e indica hasta cierto grado la respuesta final Y de la unidad experimental.
- Las covariables deben ser medidas en escala de razón, intervalar o nominal.
- Las covariables no deben estar relacionadas con los tratamientos o condiciones experimentales.
- Las covariables deben estar relacionadas con la respuesta o variable dependiente.
- No debe haber interacción entre las covariables y los tratamientos.
- La covariable debe poseer una relación lineal con la variable respuesta. De no ser así se debe aplicar una transformación para convertir la relación en lineal. Estas transformaciones pueden ser: logarítmica, recíproca, arcoseno, de raíz o exponencial.

Supuestos del ANCOVA

- Supuesto de linealidad:

La relación entre X e Y debe ser lineal. Este supuesto se prueba corriendo un análisis de regresión lineal simple entre la covariable (X) y la respuesta (Y). La regresión debe dar significativa, es decir, con valor P<0,05.

- Supuesto de homogeneidad de las pendientes

Las pendientes de todos los grupos de tratamientos deben ser iguales o aproximadamente iguales.

Este supuesto se prueba verificando que la interacción tratamiento por la covariable no sea significativa.

- Supuesto de independencia entre la covariable (X) y el tratamiento :

Este supuesto se prueba realizando un ANOVA donde la covariable (X) sea declarada como variable dependiente y los tratamientos sean declarados como variables independientes. La F de Fisher-Snedecor de este ANOVA debe resultar no significativa.

Los objetivos del análisis de covarianza son:

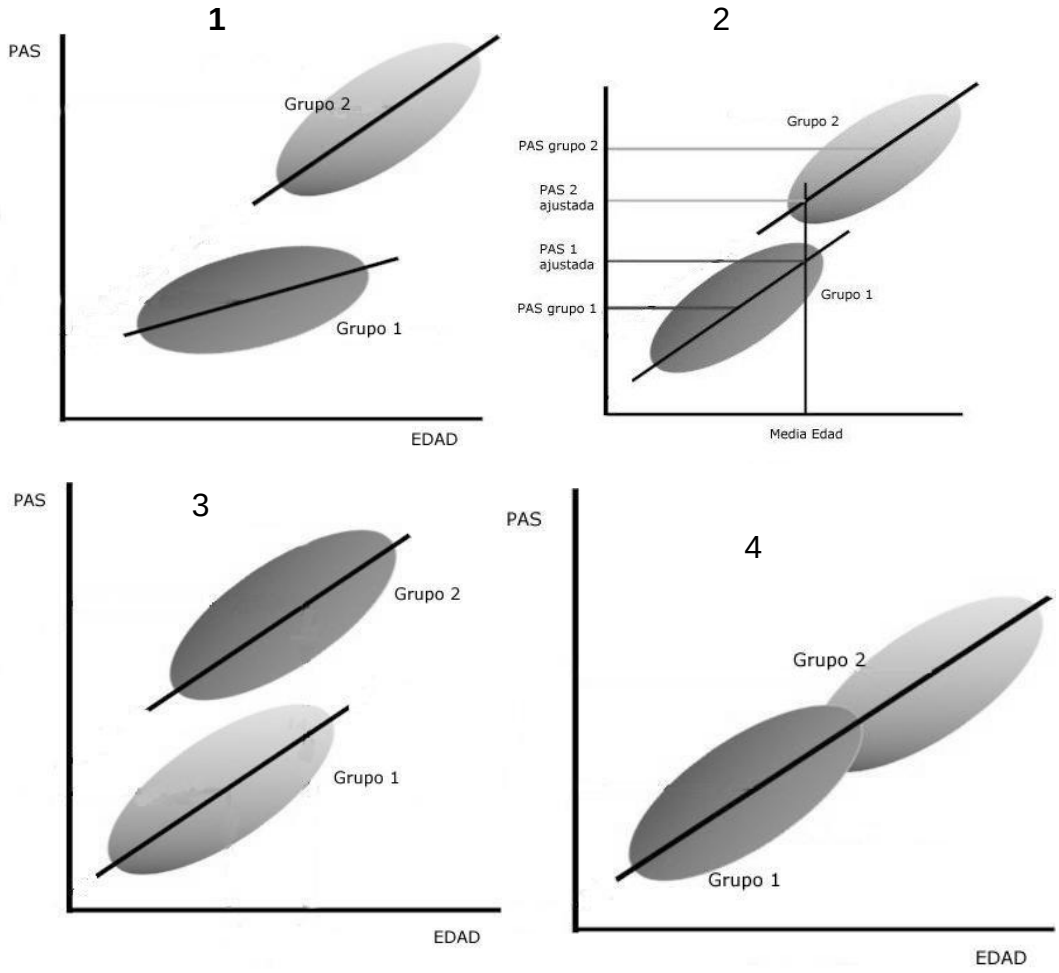
- Disminuir el error experimental con el consecuente aumento en la precisión del experimento. Lo cual puede ser corroborado a través de la disminución que experimenta el coeficiente de variación.
- Ajustar los promedios de los tratamientos. Esto puede ser corroborado comparando los promedios antes y después del ajuste. Los promedios corregidos por las covariables se llaman

Medias Mínimo Cuadráticas Corregidas (en inglés: LSMEANS), y se calculan así:

$$\hat{\overline{Y}} = \overline{Y} - b_{yx}(X_i - \overline{X})$$

- ❖ Facilitar la interpretación de los resultados en el experimento, especialmente en lo relacionado con la naturaleza de los efectos de los tratamientos.
- ❖ Estimar el valor de las unidades perdidas en los experimentos.

Veamos con un ejemplo gráfico que pretende estudiar el efecto de la presión arterial sistólica (PAS) en dos grupos de pacientes según edad (covariable):



En el gráfico 1: se aprecian dos grupos con pendientes desiguales

en esta circunstancia no se puede realizar un análisis de covarianza, porque los diversos grupos de tratamiento que se comparan deben tener pendientes iguales o aproximadamente iguales.

En el gráfico 2: se ha producido el ajuste por la covariable y las pendientes de los dos grupos son paralelas (no hay interacción)

En el gráfico 3: se aprecia claramente la forma de homogenizar los dos grupos en una sola pendiente que sirva para realizar el ajuste de los promedios de los dos grupos de tratamientos.

En el gráfico 4: se han unificado los dos grupos en una sola pendiente para llevar a cabo el ajuste de las medias de tratamiento.

Ejemplo de Aplicación de ANCOVA

Una investigación sociológica realizada en Caracas deseaba determinar la relación entre la ausencia del padre en el hogar (tratamientos) y el nivel de autoestima alcanzado por los hijos varones (Y=variable dependiente), pero se sospecha que esta relación pudiera estar afectada por el clima familiar (covariable):

Matriz de Datos

Familias con Padrastro		Familias con Padre		Familias sin Padre	
Y (autoestima)	X (clima)	Y (autoestima)	X (clima)	Y (autoestima)	X (clima)
15	30	25	28	5	10
10	20	10	12	10	15
5	15	15	20	20	20
10	20	15	10	5	10
20	25	10	10	10	10

Se pide:

(a) Correr un ANCOVA

(b) Verificar los supuestos

(c) Expresar conclusiones

Para resolver esta situación organizamos la matriz de datos en el STATISTIX así:

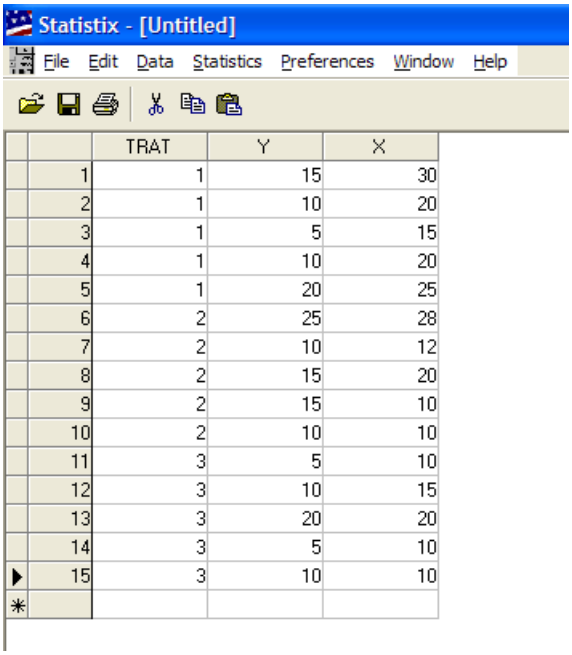
Codificamos:

Tratamiento:

1=con padrastro;

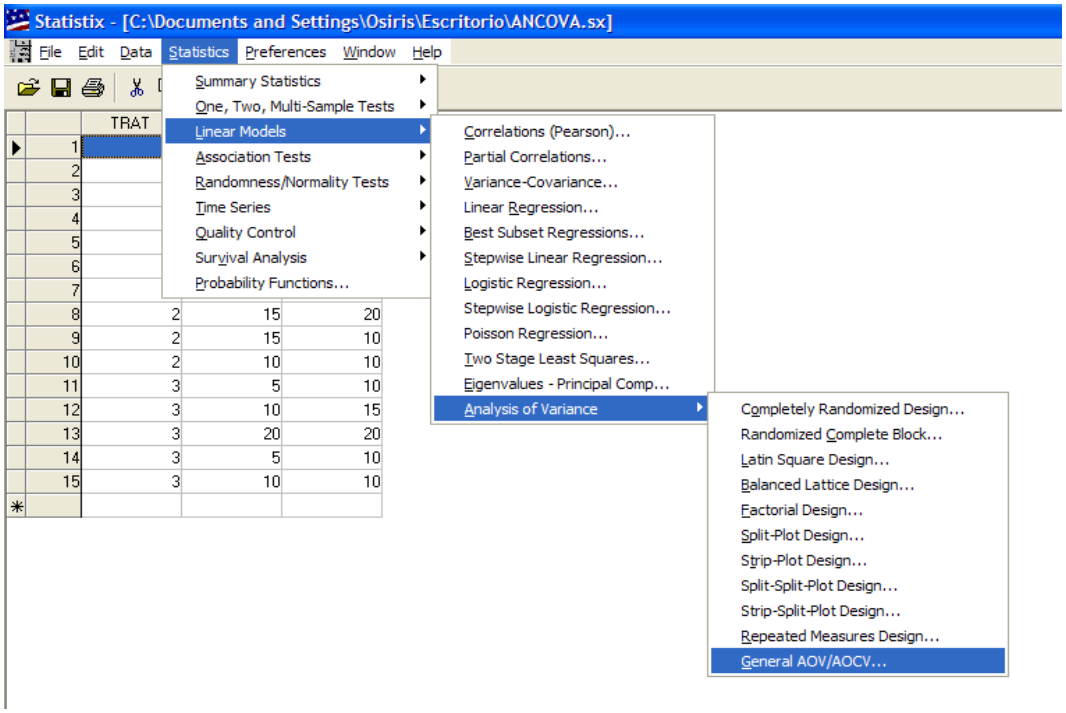
2=con padre;

3=sin padre



	TRAT	Y	X
1	1	15	30
2	1	10	20
3	1	5	15
4	1	10	20
5	1	20	25
6	2	25	28
7	2	10	12
8	2	15	20
9	2	15	10
10	2	10	10
11	3	5	10
12	3	10	15
13	3	20	20
14	3	5	10
15	3	10	10

La secuencia de opciones es:

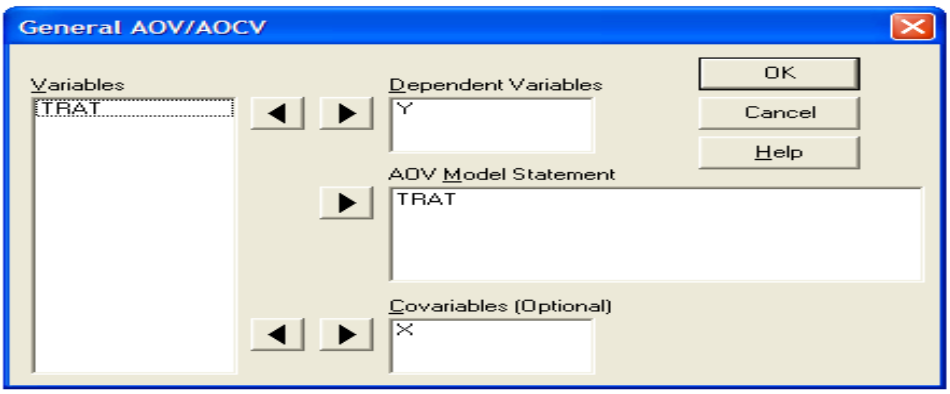


La caja de diálogo tiene tres variables que declarar:

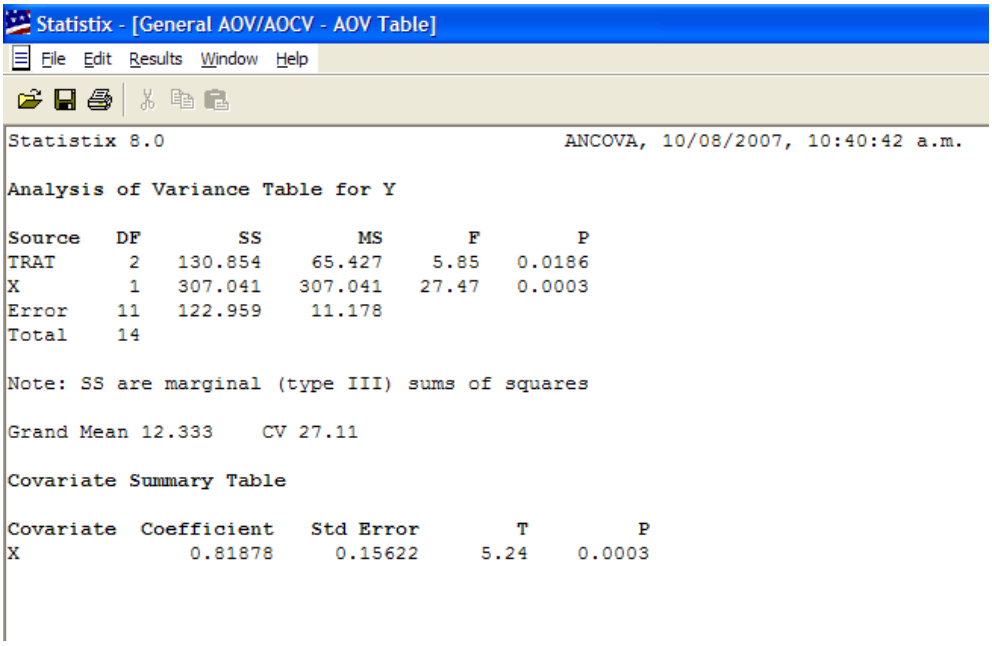
(i) La variable dependiente

(ii) La variable tratamiento

(iii) La covariable



La salida del paquete Statistix es:



Como la F de tratamiento en el análisis de la varianza global dio significativo, pasamos a verificar ¿Cuál tratamiento es el responsable

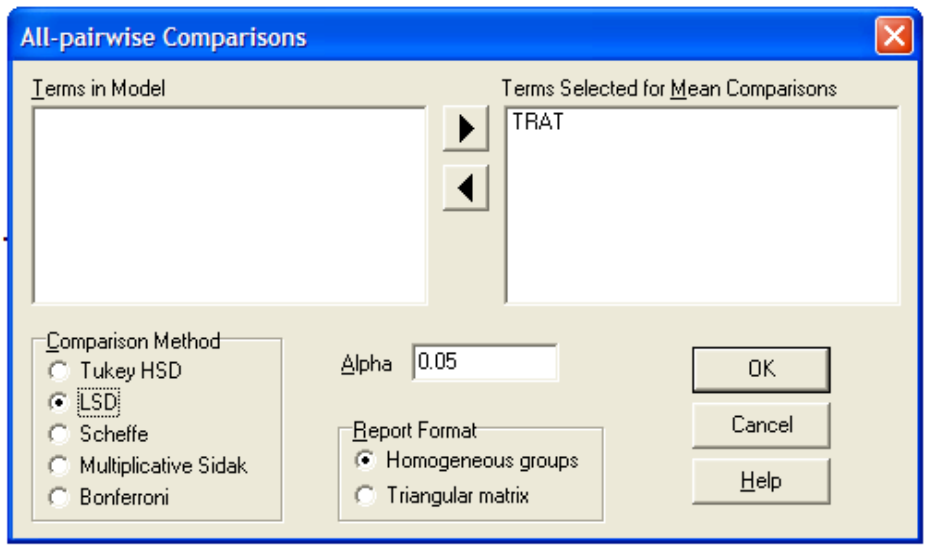
de la diferencia significativa? Esto lo podemos realizar con la prueba Mínima Diferencia Significativa así:

(a) No vamos a la opción de resultados.

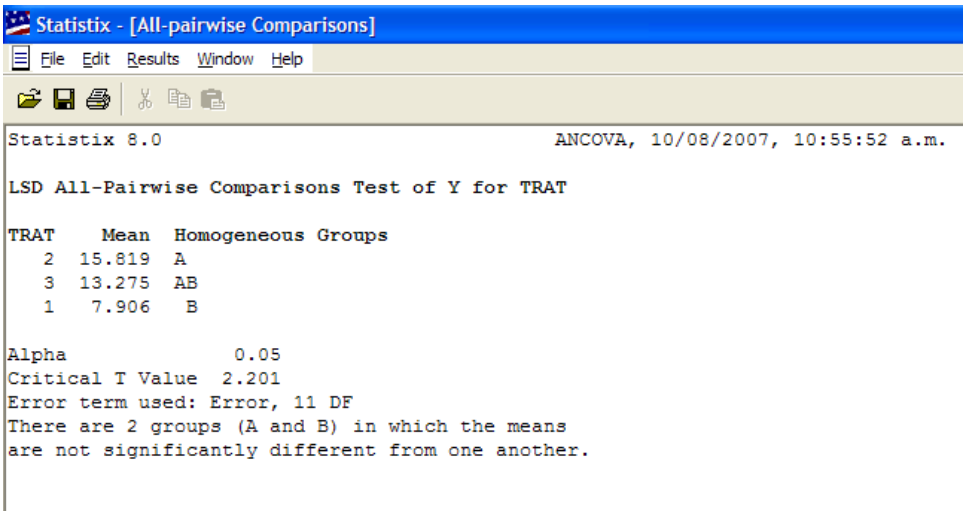
(b) Buscamos la opción Comparaciones Múltiples

(c) Todas las comparaciones por parejas

La caja de diálogo es:



La salida es:



Conclusiones

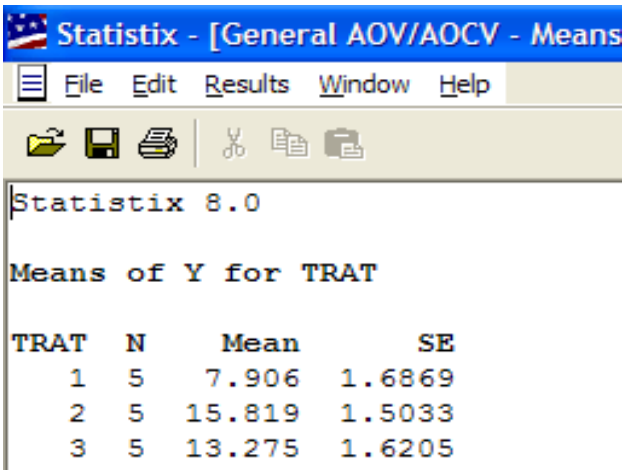
(a) Existe diferencias significativas entre los tratamientos F=5,86; P=0,0186. Lo cual indica que la autoestima difiere por la variable tratamiento (presencia del padre)

(b) La prueba de LSD revela que la autoestima es mayor en el grupo 2 (con presencia del padre) y no presenta diferencia significativa con el grupo 3 (sin padre). El menor puntaje de autoestima se presentó en el grupo 1 (con padrastro).

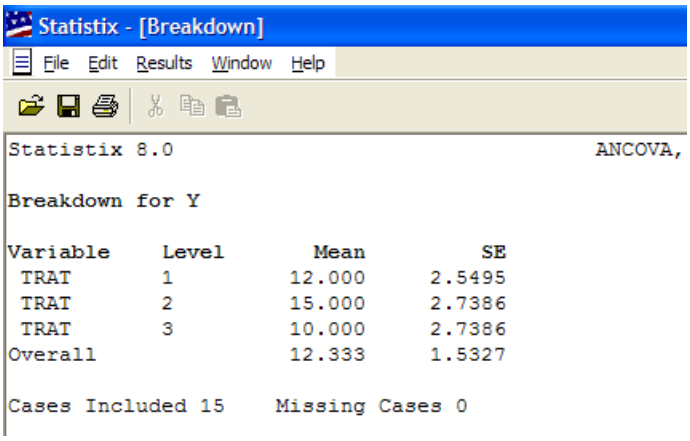
(c) La covariable (Clima familiar) es significativa t=5,24; P=0,0003. Lo cual indica que esta variable influye en la autoestima.

Los promedios de autoestima no corregidos y corregidos se dan a continuación:

Promedios Corregidos de Autoestima:



Promedios No Corregidos de autoestima:



Es evidente el ajuste del nivel de autoestima por la covariable (clima familiar) por dos razones:

(i) Menor promedio ajustado en los valores corregidos como es de esperar.

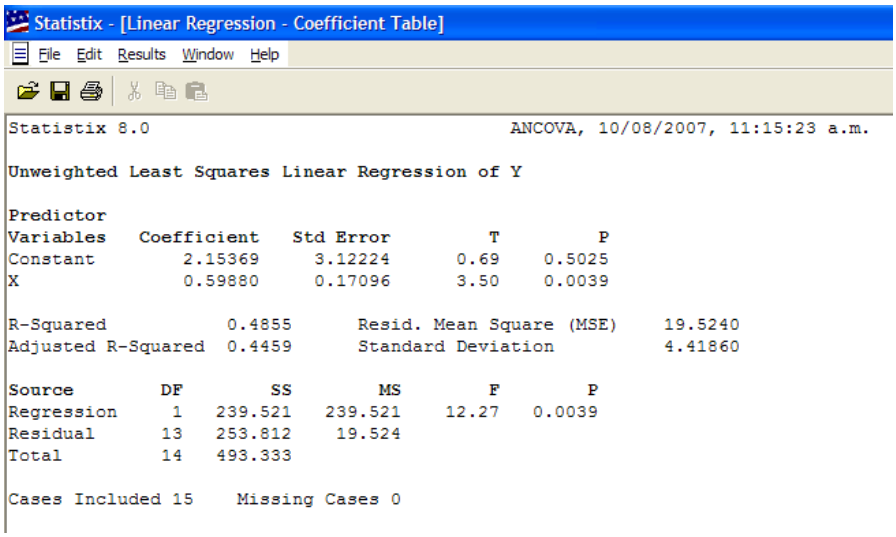
(ii) Menores errores estándar por promedio ajustado como es de esperar.

Prueba de los Supuestos:

(1) Prueba de Linealidad

Ho: la relación X e Y no es lineal ($\beta = 0$)

H1: La relación es lineal ($\beta \neq 0$)



Statistix - [Linear Regression - Coefficient Table]

File Edit Results Window Help

```
Statistix 8.0                                ANCOVA, 10/08/2007, 11:15:23 a.m.

Unweighted Least Squares Linear Regression of Y

Predictor
Variables   Coefficient   Std Error         T         P
Constant        2.15369     3.12224      0.69    0.5025
X               0.59880     0.17096      3.50    0.0039

R-Squared           0.4855      Resid. Mean Square (MSE)     19.5240
Adjusted R-Squared  0.4459      Standard Deviation           4.41860

Source        DF         SS         MS         F         P
Regression     1    239.521    239.521     12.27    0.0039
Residual      13    253.812     19.524
Total         14    493.333

Cases Included 15    Missing Cases 0
```

Luego se cumple el primer supuesto del ANCOVA, aceptamos en este caso H1: la relación es lineal.

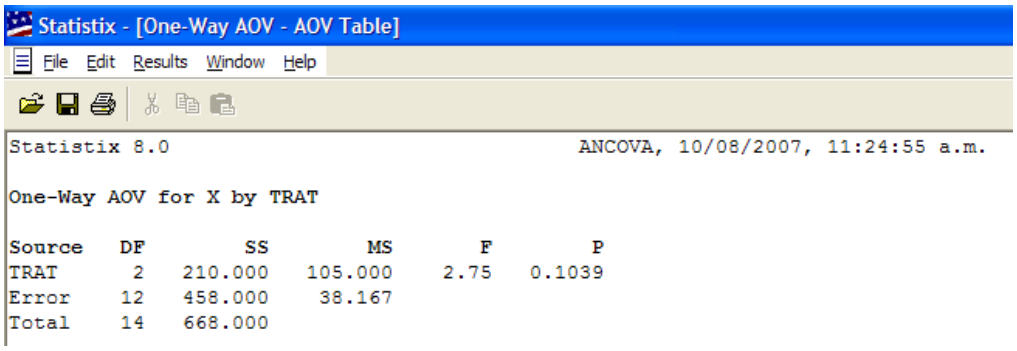
(2) Supuesto de independencia entre la covariable y el tratamiento

La Hipótesis es:

Ho: La covariable es independiente de los tratamientos

Ho: La covariable es dependiente de los tratamientos

Con un ANOVA así:



No se rechaza Ho (F=2,75; P=0,1039). Lo cual indica que existe independencia entre la covariable y el tratamiento.

Luego se cumple el supuesto de independencia.

(3) Supuesto de interacción entre la covariable y el tratamiento

Ho: No hay interacción entre X(covariable) y los tratamientos.

Ho: Hay interacción entre X(covariable) y los tratamientos.

La prueba de esta interacción no puede procesarse en el Statistix porque por ahora no dispone del algoritmo apropiado.

La prueba realizada en el SPSS, da un valor F=2,87; P=0,169 para la interacción. Lo cual indica que no se puede rechazar Ho. En consecuencia, no hay interacción entre la covariable y el tratamiento. Se cumple el supuesto de interacción.

La prueba de los supuestos debe realizarse antes de corre el análisis de covarianza propiamente dicho.

El incumplimiento de uno cualquiera de los tres supuestos invalida el análisis de covarianza.

Capítulo VII

Análisis de Varianza no paramétricos

Anova de Kruskal-Wallis
Anova de Friedman
Anova de Q de Cochran

Introducción

Las técnicas de análisis de varianza no paramétricos son útiles cuando los supuestos de: Normalidad, Homogeneidad de las varianzas, Independencia de los Errores y Aditividad de los efectos no se cumplan.

Las pruebas para verificar el cumplimiento de estos supuestos se resumen en el cuadro siguiente:

Supuesto	Prueba	Criterio
Normalidad	Kolmogorov-Smirnov Shapiro-Wilk	Comparación de Distribuciones W de Shapiro-WilK no significativo
Aditividad de	Prueba de Aditividad	Aceptación de Ho:Los efectos

los efectos	de Tukey	son aditivos
Homogeneidad De Varianzas	Prueba de Levene Prueba de Bartlett	Aceptación de Ho: las varianzas son iguales
Independencia de los errores	Prueba de la rachas ó Prueba de aleatoriedad	Aceptación de Ho: los efectos son independientes

Técnicas No Paramétricas según Tipo de Escala de Medición

Las técnicas no paramétricas se pueden agrupar así:

Nivel de Medida	Nominal	Ordinal
Una muestra	Binomial Ji-Cuadrado	Kolmogorov-Smirnov Rachas
Dos Muestras Relacionadas	McNemar	Signos Wilcoxon
Dos muestras independiente	Fisher Ji-Cuadrado	Mediana U de Mann-Whitney Kolmogorov-Smirnov
K Muestras relacionadas	Q de Cochran	Friedman
K Muestras independientes	Ji-cuadrado	Kruskal-Wallis

En este capítulo sólo desarrollaremos las Pruebas para K muestras independientes y dependientes.

Técnicas Paramétricas Para K Muestras Independientes

Prueba de Kruskal-Wallis: Análisis de varianza por rangos

El análisis de variancia en un sentido por rangos de Kruskal-Wallis compara tres o más muestras para definir si provienen de poblaciones iguales.

Requerimientos:

- Se requiere la escala ordinal de medición.
- Es una alternativa para ANOVA en un sentido.
- La distribución Ji-cuadrado es el estadístico de prueba.
- Cada muestra debe tener al menos cinco observaciones.
- Los datos de la muestra se jerarquizan de menor a mayor como si fueran de un solo grupo.

El estadístico de Prueba está dado por:

$$H = \frac{12}{n(n+1)}\left\{\frac{(\Sigma R_1)^2}{n_1} + \frac{(\Sigma R_2)^2}{n_2} + \ldots + \frac{(\Sigma R_k)^2}{n_k}\right\} - 3(n+1)$$

Ejemplo de Aplicación

Keely Ambrose, director de recursos humanos, estudia el porcentaje de aumento en el salario de la gerencia media en cuatro de sus plantas manufactureras. Obtuvo una muestra de gerentes y determinó el porcentaje de aumento en su salario. Para 5% de nivel de significancia ¿puede Keely concluir que hay una diferencia en el porcentaje de aumento?

Milville	Rango	Camden	Rango	Eaton	Rango	White Plains	Rango
2.2	2	1.9	1	3.7	6	5.7	9
3.6	5	2.7	3	4.5	7	6.8	10.5
4.9	8	3.1	4	7.1	13.5	8.9	16
6.8	10.5	6.9	12	9.3	17	11.6	18.5
7.1	13.5	8.3	15	11.6	18.5	13.9	20
Total	39		35		62		74

- Paso 1: H_0: las poblaciones son iguales.
- H_1: las poblaciones no son iguales.
- Paso 2: H_0 se rechaza si $x^2 > 7.185, gl = 3, \alpha = .05$
- Paso 3: estadístico de prueba $z = 5.95$
- Paso 4: H_0 no se rechaza. No hay diferencia en las poblaciones.

El análisis de la varianza de Friedman

Cuando K muestras igualadas tienen sus observaciones medidas, por lo menos, en la escala ordinal, el análisis de la varianza de dos criterios de Friedman puede ser utilizado para probar si las K muestras han sido obtenidas de poblaciones diferentes.

El arreglo en bloques consiste en colocar los datos en una tabla de doble entrada de n filas y k columnas. Las filas (bloques) representan a los distintos sujetos, unidades, animales, plantas, etc, etc., y las columnas a las diferentes condiciones (tratamientos, grupos, muestras, etc.)

Al obtener los datos, éstos deben ser ordenados por rangos de 1 a K.

Para cada condición (tratamiento) asumamos los rangos y denominamos este total R_j para la j-ésima columna.

Para la Prueba de Friedman usamos el estadístico Ji-cuadrado dado por:

$$\chi_r^2 = \frac{12}{nk(k+1)} \sum R_{.j}^2 - 3(K+1)$$

Donde:
N: número de filas o bloques
K: número de tratamientos
R_j : es la suma de los rangos de la j-ésima columna.

Conforme aumenta la cantidad de bloques en el experimento (más de 5) se puede aproximar el estadístico de Friedman a una distribución χ^2 con (k-1) grados de libertad.

Ejemplo de Aplicación del ANOVA de Friedman

Se diseña un experimento de pruebas de degustación de modo que cuatro marcas de café colombiano sean clasificados por 9 expertos.

Para evitar cualquier efecto acumulado, la sucesión de pruebas para las 4 infusiones se determina aleatoriamente para cada uno de los 9 probadores expertos hasta que se dé una clasificación en una escala de 7 puntos (1=en extremo desagradable, 7= en extremo agradable

para cada una de las siguientes 4 categorías: sabor, aroma, cuerpo y acidez) la suma de los puntajes de las 4 característica.

Los datos para cada experto se convierten a rangos.

El sistema hipotético

Ho: las medianas de los resultados sumados (para las cuatro características) son iguales.

H1: Por lo menos dos marcas tengan resultados diferentes.

Ho: Md1=Md2=Md3=Md4 (las medianas son iguales)

H1: por lo menos dos de las medianas son diferentes

Marcas				
Experto	**A**	**B**	**C**	**D**
1	**24**	**26**	**25**	**22**
2	**27**	**27**	**26**	**24**
3	**19**	**22**	**20**	**16**
4	**24**	**27**	**25**	**23**
5	**22**	**25**	**22**	**21**
6	**26**	**27**	**24**	**24**
7	**27**	**26**	**22**	**23**
8	**25**	**27**	**24**	**21**
9	**22**	**23**	**20**	**19**

La conversión de esta matriz de datos en Rangos (por filas) es:

Marcas				
Experto	**A**	**B**	**C**	**D**
1	**2.0**	**4.0**	**3.0**	**1.0**
2	**3.5**	**3.5**	**2.0**	**1.0**
3	**2.0**	**4.0**	**3.0**	**1.0**
4	**2.0**	**4.0**	**3.0**	**1.0**
5	**2.5**	**4.0**	**2.5**	**1.0**
6	**3.0**	**4.0**	**1.5**	**1.5**
7	**4.0**	**3.0**	**1.0**	**2.0**
8	**3.0**	**4.0**	**2.0**	**1.0**
9	**3.0**	**4.0**	**2.0**	**1.0**
Media Rango	**25**	**34,5**	**20**	**10,5**

$$\chi_r^2 = \frac{12}{nk(k+1)} \sum R_{.j}^2 - 3(K+1)$$

$$\chi_r^2 = \left[\frac{12}{9(4)(4-1)} \left(25^2 + 34.5^2 + 20^2 + 10.5^2\right) \right] - 3(9)(4+1) = 20.03$$

$$\chi_r^2 = \frac{12}{9x4x5} \left(25^2 + 34.5^2 + 20^2 + 10{,}5^2\right) - 3(9)(5) = 20{,}03$$

Puesto que F es mayor que el valor tabulado por tanto se rechaza Ho.

Se puede concluir que hay diferencias importantes (percibidas por los expertos) con respecto a la calidad de las 4 marcas de café.

Una vez rechazado Ho la hipótesis nula se pueden usar técnicas de comparaciones múltiples a posteriori para determinar qué grupo o grupos, difieren significativamente de los demás. Dada la magnitud de las medias se sugiere la Prueba de Mínima Diferencia Significativa.

Prueba Q de Cochran

Análisis de la Varianza de dos vías sin interacción con respuesta dicotómica (Binaria)

Frecuentemente diseñamos experimentos de tal manera que más de dos muestras o condiciones pueden estudiarse simultáneamente.
La Q de Cochran es una prueba para comparar las proporciones de respuestas de un tipo (positivo o negativo) o (cero o uno) de varios sujetos bajo ciertas condiciones de tratamiento.

Es una prueba para K muestras relacionadas porque los mismos sujetos son evaluados bajo las mismas condiciones de tratamiento.

Matriz de Datos

Es un arreglo de datos binarios en tablas de doble entrada con n sujetos y K condiciones de tratamiento.

Hipótesis

Ho: P1=P2=P3=…=Pk

H1: Algún Pi es diferente

El estadístico Q de Cochran sigue una distribución Ji-Cuadrado con K-1 grados de libertad.

La prueba Q de Cochran es una generalización de la Prueba de McNemar.

El estadístico Q de Cochran se calcula con la siguiente fórmula:

$$Q = \frac{(K-1)\left[K\sum G_i^2 - (\sum G_i)^2\right]}{K\sum H_j - \sum H_j^2}$$

Donde:
G= total de la i-ésima columna (condición experimental)
H= total de la j-ésima fila o hilera (sujeto)

Ejemplo de Aplicación

Se desean comparar tres métodos de diagnóstico para la brucelosis bovina (M1,M2 y M3) para ello se tomaron al azar 14 sueros bovinos y se determino por cada método su positividad (resultado uno) y no positividad (resultado cero):

Suero #	Método 1	Método 2	Método 3
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	1	0	0
6	1	0	0
7	1	0	0
8	1	1	0
9	1	1	0
10	1	1	0
11	1	1	0
12	1	1	1
13	1	1	1
14	1	1	1

Se pide:

Verificar si los tres métodos de diagnóstico son iguales o diferentes en su especificidad. De ser diferentes indicar ¿Cuál es el mejor?

Hipótesis de Investigación

Las respuestas son iguales con los tres métodos

Hipótesis Estadísticas

Ho: P1=P2=P3=…=Pk

H1: algún Pi es diferente

Suero #	Método 1	Método 2	Método 3	Hj	H_j^2
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	1	0	0	1	1
6	1	0	0	1	1
7	1	0	0	1	1
8	1	1	0	2	4
9	1	1	0	2	4
10	1	1	0	2	4
11	1	1	0	2	4
12	1	1	1	3	9
13	1	1	1	3	9
14	1	1	1	3	9
Gj	10	7	3	$\sum H_j = 20$	$\sum H_j^2 = 46$
G_j^2	100	49	9	$\sum G_i^2 = 158$	$\sum G_i = 20$

Calculemos Q:

$$Q = \frac{(K-1)\left[K\sum G_i^2 - (\sum G_i)^2\right]}{K\sum H_j - \sum H_j^2}$$

$$Q = \frac{[3-1][3(158) - (20)^2]}{3(20) - 46} = 10{,}57$$

Decisión: Como Q=10,57 es mayor que el valor tabulado de Ji-cuadrado = 9,21, rechazamos Ho.

Conclusión: las respuestas no son todas iguales con los tres métodos.

El mejor método es el uno. Que puede verificarse con una prueba a posteriori de mínima diferencia significativa.

Capítulo VIII

Análisis de Correlación Simple, Múltiple, Parcial

Correlación

Es la medida del grado de relación entre dos o más variables.
Con variables nominales suele utilizarse el término *Asociación* para indicar el grado de relación entre las variables.

Correlación Simple

La correlación entre dos variables cuantitativas para verificar su relación se llama: **Correlación Simple**, porque sólo involucra una variable independiente. Mientras que la relación entre varias variables independientes con una dependiente se le llama: **Correlación Múltiple**.

La relación entre dos variables manteniendo el resto constante recibe el nombre de **Correlación Parcial**.

La correlación con una sola variable independiente se llama: Simple.

La correlación con más de una sola variable independiente se llama: Múltiple.

La correlación de un grupo de variables dependientes con un grupo de variables independientes, es decir, entre grupos de variables se llama: **Correlación Canónica.**

Coeficiente de Correlación según la naturaleza de las variables

El grado de relación entre variables depende de la naturaleza de las variables involucradas en el estudio o investigación.
En este sentido, si ambas variables son nominales la relación será descrita con el Estadístico Ji-Cuadrado. Si ambas variables son

ordinales se describe la relación con el Coeficiente de Correlación de Spearman. Si ambas variables son intercalares mediante el Coeficiente de Pearson. Si una variable es nominal y la otra es intervalar la relación puede ser descrita mediante el Coeficiente Omega Cuadrado. Si ambas variables son dicotómicas o binarias la relación puede establecerse mediante el Coeficiente Phi.

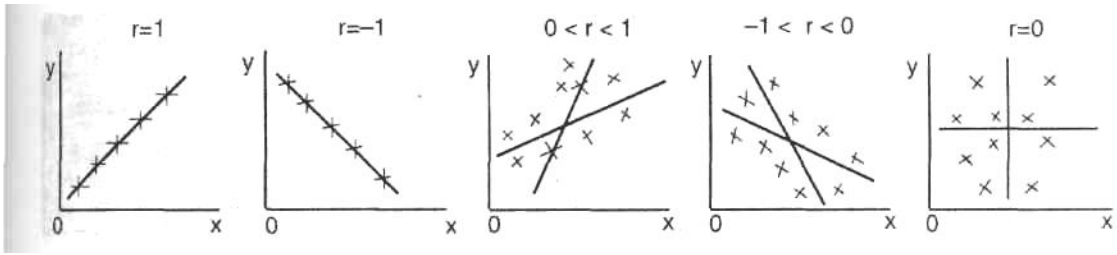
La relación entre dos fenómenos puede ser: estricta, funcional o nula. La relación entre talla y peso no es estricta, ya que no existe una proporcionalidad simple entre ambas variables; pero tampoco es nula, pues de lo contrario ambas variables serían independientes entre sí (mientras que en este caso existe, indudablemente, cierta correlación: la gente alta es, en general, más pesada). No obstante, esta relación no resulta funcional, porque de lo contrario se podría captar matemáticamente la dependencia mediante una ecuación de dos variables.

La correlación tiene las mismas propiedades de los vectores: magnitud, dirección y sentido.

En tal sentido, se habla de Correlación Positiva o Directa cuando ambas características (expresadas mediante valores de las variables) presentan la misma tendencia (por ejemplo, talla y peso) porque a medida que una aumenta se espera que la otra variable aumente pero esta relación en los seres vivos no es indefinida sino hasta cierta edad. Es decir, hay una variable reguladora la edad en este caso.

En este mismo orden de ideas, se habla de Correlación Negativa o Inversa cuando una variable aumenta y la otra disminuye, mostrando tendencias claramente opuestas, es el caso de la oferta y el precio en el ámbito de la economía. Cuando la oferta aumenta, el precio tiende a bajar.

Los diagramas de dispersión o Scattergramas suelen ser útiles para estudiar el grado de relación entre dos variables.



Positiva Perfecta | Negativa perfecta | Aleatoria positiva | Aleatoria negativa | Nula

Correlación como medida de la Confiabilidad de un Instrumento de medición o Test

La correlación es la base utilizada para evaluar la confiabilidad de un instrumento de medición o test. Si los puntajes de un test fueron medidos en base a una escala Likert o tipo Likert, se utilizará el **Coeficiente Cronbach**, pero si los puntajes provienen de alternativas dicotómicas o binarias (si, no) se utilizará el **Coeficiente de Kuder-Richardson**. Una interrogante que salta a la mente de inmediato es ¿Cómo analizar un Test que tiene preguntas en escala (si, no) y en escala Likert? Por supuesto que esta situación nos conduce a trabajar más en el sentido que debe aplicarse el test en dos oportunidades diferentes y luego correlacionar mediante un **coeficiente de correlación de Pearson**, la relación de los puntajes totales de la primera aplicación con los de la segunda. Si se mantiene entre las dos aplicaciones una correlación que por lo menos es mayor que 0,70 se concluye que el test es confiable.

Por otra parte, existe un coeficiente de partición por mitades o **Correlación de Spearman-Brown** que mide el grado de homogeneidad de un test; cuando las correlaciones entre la primera y la segunda mitad del test, o entre pares e impares es lo más elevada posible y en todo caso mayor que 0,70, se concluye igualmente que el test es confiable.

La Correlación también hace posible el cálculo del **Coeficiente de Determinación** R^2 que se utiliza como medida de la **Bondad de Ajuste de un Modelo de Regresión**, como se verá más adelante.

En general, si el valor de R-cuadrado es mayor en comparación a otro modelo, el modelo que posea un R-cuadrado mayor será el de mejor ajuste. El R-cuadrado, comienza a ser importante si sobrepasa el valor 0,70. Este coeficiente siempre es positivo.

Ejemplos de Aplicación de la Correlación

Correlación de Pearson

Correlación Producto Momento conocida también como Correlación de Pearson o Correlación de Bravais-Pearson.

El STATISTIX contiene en su algoritmo computacional la fórmula:

$$r_{xy} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}$$

Como se observa en el numerador tenemos la fórmula de covarianza (X,Y) y en el denominador la raiz cuadrada de la varianza de X y Y.

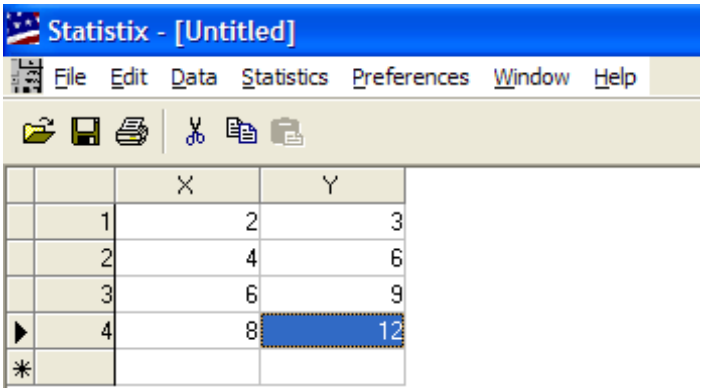
Una vez vaciados los datos en el paquete informático de computación se sigue la secuencia:

STATISTIX>LINEAR MODELS>CORRELATION PEARSON

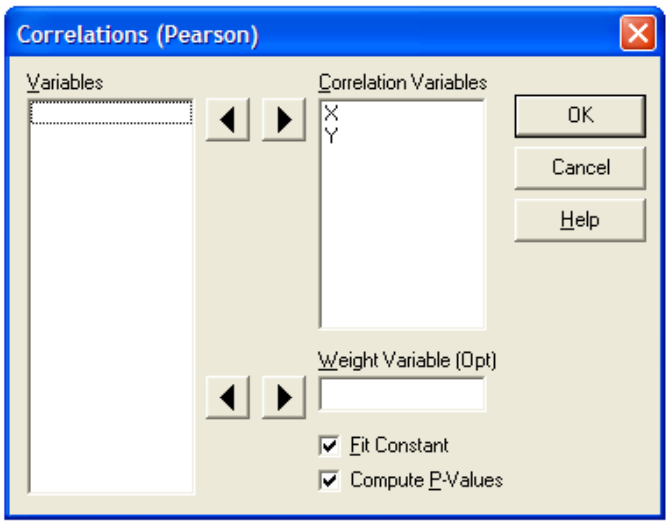
Sea la determinación de la correlación simple entre X e Y:

X	2, 4, 6, 8
Y	3, 6, 9, 12

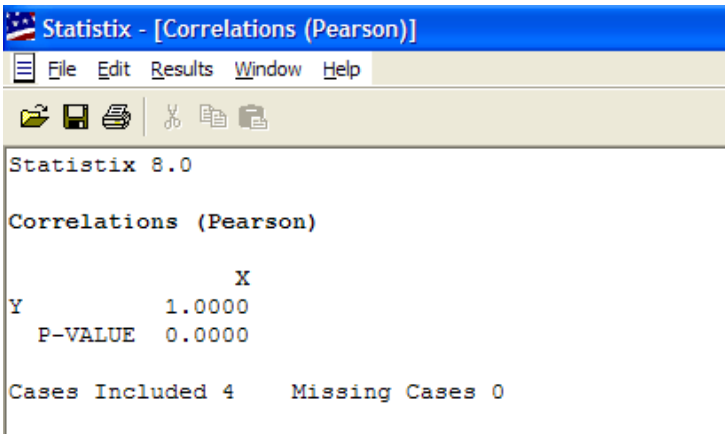
Una vez introducidos los datos,



Seguimos la secuencia dada anteriormente:



Cuya salida es:



El resultado anterior revela una correlación alta, directa y perfecta.

Alta porque da mayor de 0,70, directa porque al crecer, crece Y en la misma proporción (lo indica el signo positivo del coeficiente de correlación de Pearson) y perfecta porque dio uno.

Existe una escala para interpretar el coeficiente de correlación dada por algunos autores:

Rango	Significado
0,00 a 0,29	Bajo
0,30 a 0,69	Moderado
0,70 a 1,00	Alto

Sin embargo, no existe un acuerdo entre los distintos autores, entre otros (Hernández, 2003, p.532) y encontramos otros baremos de interpretación:

Magnitud de la Correlación	Significado
-1,00	Correlación negativa perfecta
-0,90	Correlación negativa fuerte
-0,75	Correlación negativa considerable
-0,50	Correlación negativa media
-0,10	Correlación negativa débil
0,00	Correlación nula
+0,10	Correlación positiva débil
+0,50	Correlación positiva media
+0,75	Correlación positiva considerable
+0,90	Correlación positiva muy fuerte
+1,00	Correlación positiva perfecta

Valor P de Significación de R

Otro aspecto importante a considerar es la significancia del valor de r, que viene dado por el valor P que lo acompaña.

Si el valor P que acompaña a R es menor que 0,05, concluímos que la correlación es significativa y esto indica que es una correlación o relación real, no debida al azar. Por ejemplo, si la salida del software muestra un R=0,80; P<0,05, nos indica que la correlación es significativa.

Varianza de factores Comunes

El valor de R elevado al cuadrado indica el porcentaje de la variación de una variable debida a la variación de la otra y viceversa.

Por ejemplo, si la correlación entre productividad y asistencia al trabajo es de 0,80. Esto es,

$$r = 0{,}80$$

$$r^2 = 0{,}84$$

Expresa que la productividad explica el 64% de la variación de la asistencia al trabajo o que la asistencia al trabajo explica el 64% de la productividad.

Correlación de Spearman

Esta correlación mide la relación entre dos variables ordinales.

Por ejemplo, para correlacionar los puntajes de dos test medidos en una escala Likert.

Ejemplo, sea la determinación de la correlación entre los ordenes de llegada dado por dos jueces en 8 competencias de natación.

Número de Competencias

Nadador	1	2	3	4	5	6	7	8
Juez 1	10	11	9	13	7	14	6	15
Juez 2	11	13	8	10	9	15	7	14
Diferencia (D)	-1	-2	+1	+3	-2	-1	-1	+1
D^2	1	4	1	9	4	1	1	1

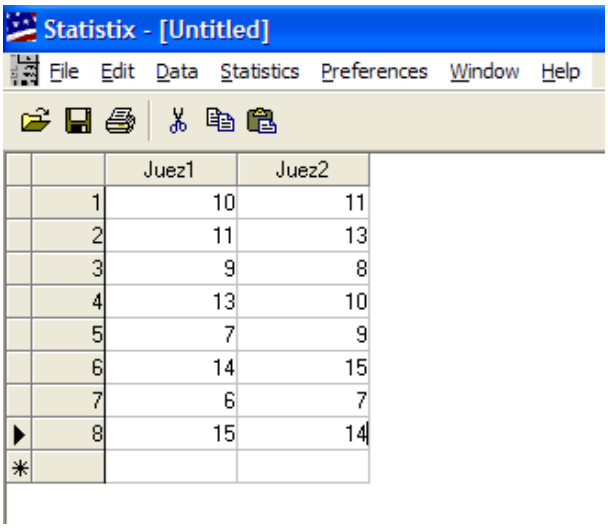
El coeficiente de Spearman se calcula así:

$$r_s = 1 - \frac{6\sum D^2}{n(n^2-1)}$$

$$r_S = 1 - \frac{6(22)^2}{8(8^2-1)} = 0{,}74$$

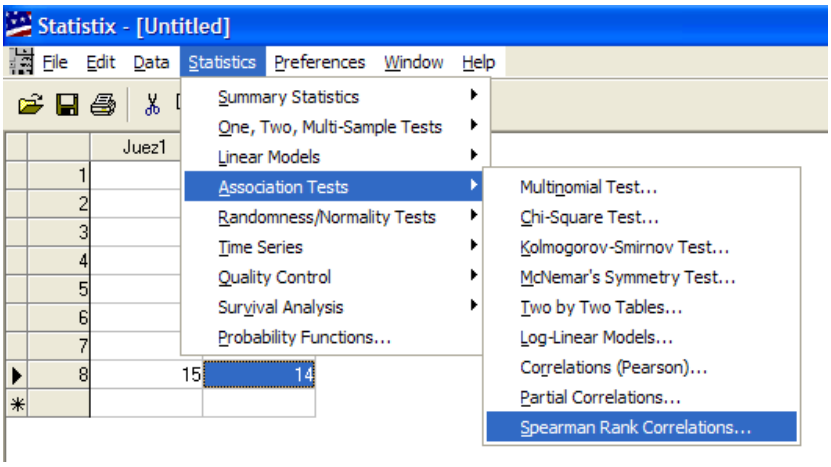
El valor de 0,74 nos indica que existe una correlación positiva relativamente alta entre la puntuación dada por uno y otro juez, es decir, que el nadador que obtuvo un puesto de llegada alto en un juez también lo obtuvo con el otro. Y así mismo, el que obtuvo baja con un juez, obtuvo baja con el otro.

Con el software vaciamos los datos:

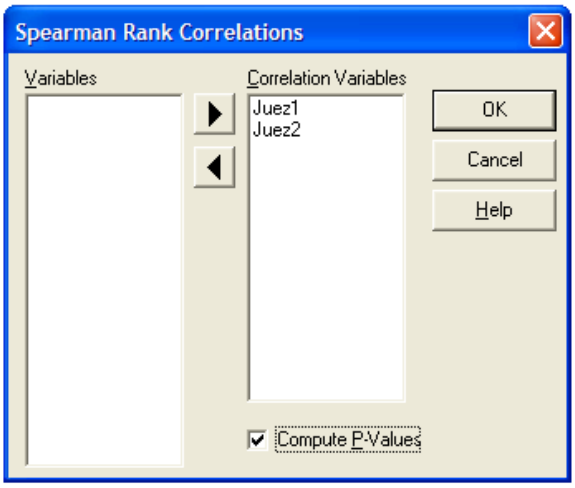


	Juez1	Juez2
1	10	11
2	11	13
3	9	8
4	13	10
5	7	9
6	14	15
7	6	7
8	15	14

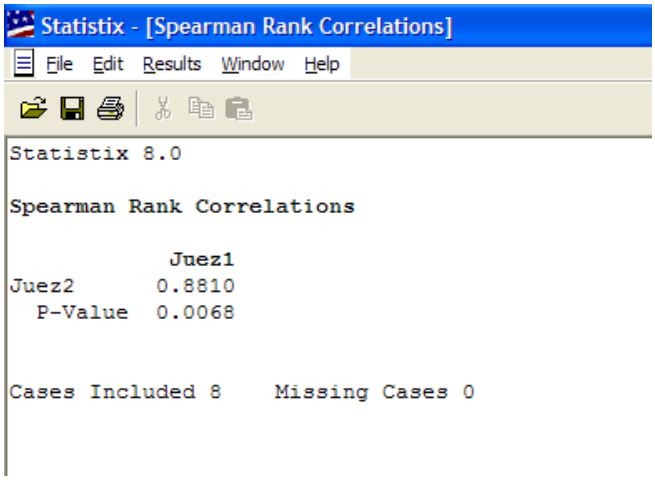
Seguimos la secuencia:



Aparece la caja de diálogo siguiente:



Mostrando la salida (output) siguiente:



La diferencia en el resultado a mano y con el software se atribuye a errores de redondeo. No obstante se interpreta de la misma manera el resultado obtenido.

Correlación entre dos variables dicotómicas

Este coeficiente de asociación que abordaremos mide la correlación entre dos variables nominales (de si y no, o de 1 y 0).

El coeficiente de asociación Phi (ϕ) se calcula de la siguiente manera.

Veamos un ejemplo, se desea encontrar la asociación entre acierto en la elección de la carrera y haber recibido o no orientación

vocacional. De una muestra de 50 estudiantes universitarios se registró el siguiente arreglo en una tabla 2x2:

	Éxito	Fracaso	Total
Orientados	19	11	30
No orientados	5	15	20
Total	24	26	50

Este arreglo en símbolos es:

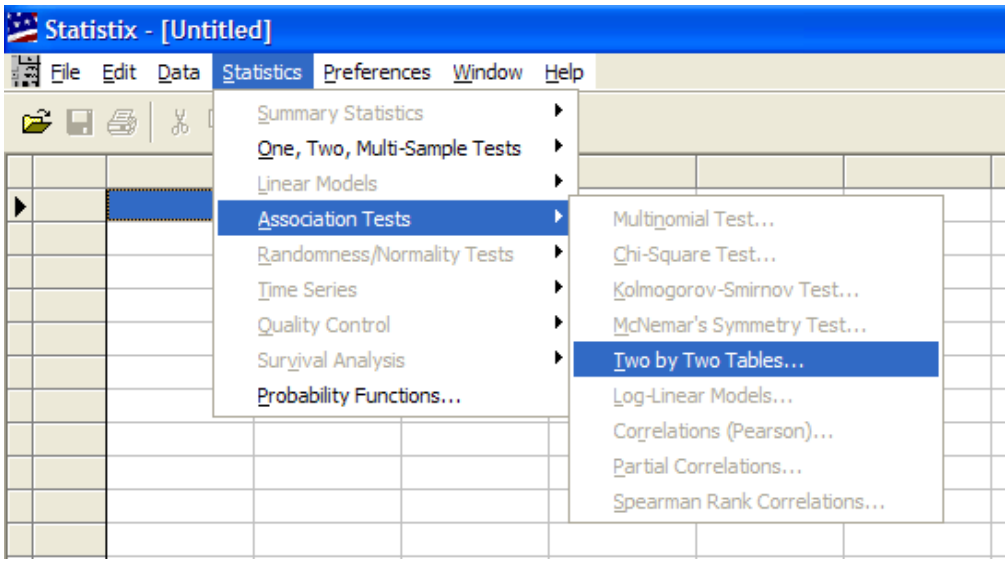
	Éxito	Fracaso	Total
Orientados	A	B	A+B
No orientados	C	D	C+D
Total	A+C	B+D	A+B+C+D

Para Phi la fórmula es:

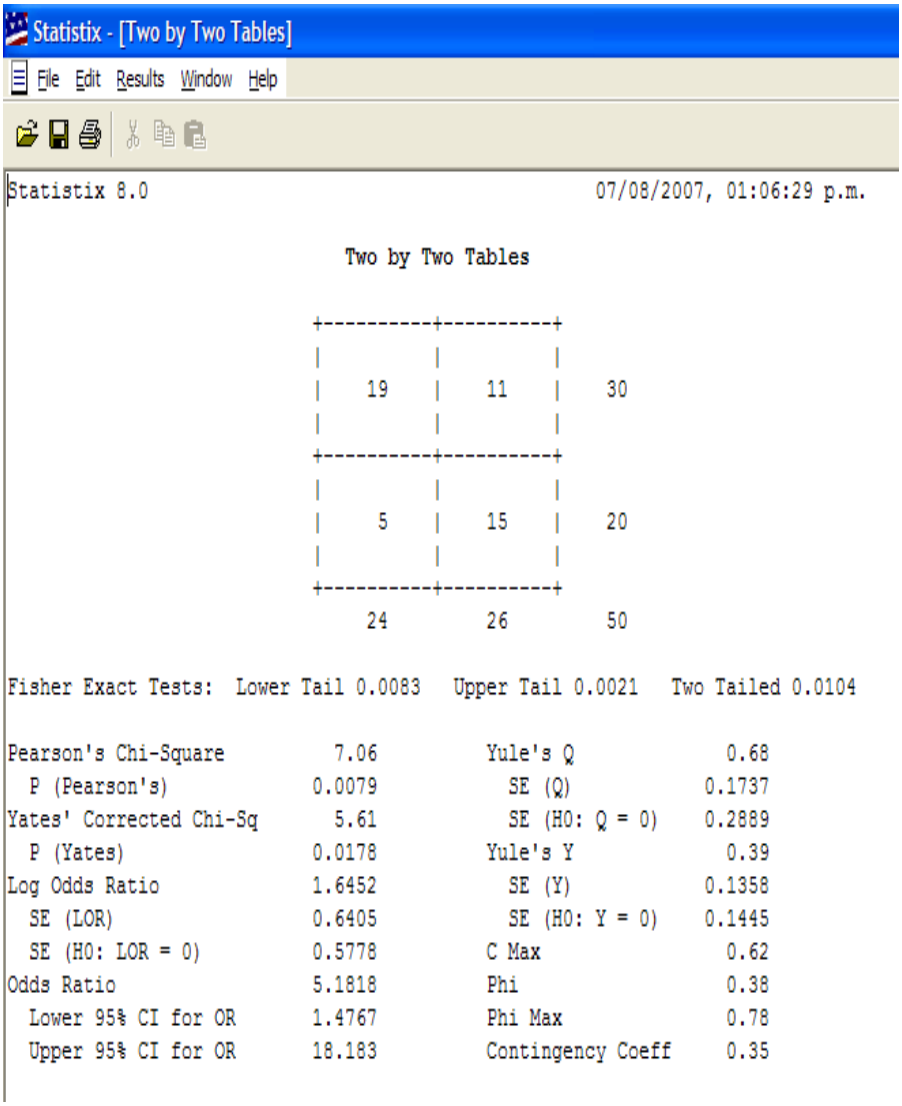
$$\phi = \frac{AxD - BxC}{\sqrt{(A+C)(B+D)(A+B)(C+D)}}$$

$$\phi = \frac{19x15 - 11x5}{\sqrt{(24)(26)(30)(20)}} = 0{,}376$$

Para interpretar Phi como se sabe que existe una relación directa entre Phi y ji-cuadrado (Phi es igual a la raíz cuadrada de Ji sobre N) Entoces podemos usar esta relación para afirmar que si ji-cuadrado es significativo también lo es Phi. En realidad Phi es una variación de la fórmula del coeficiente de correlación r de Pearson.
Veamos su determinación usando el software. Para hacerlo no necesitamos un archivo como tal, sino introducimos los datos mediante el teclado. Así:



La opción Two by Two Tables no da el siguiente cuadro de diálogo:

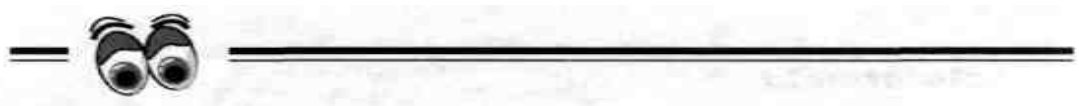


De lo antes expuesto, apreciamos que Phi=0,38 y es significativo si inspeccionamos el valor correspondiente a Ji-cuadrado

(Ji=7,06;P=0,0079) vemos que también Phi es significativo con un P=0,0079.

Correlación Biserial Puntual

Es una medida de la relación que puede haber entre una variable continua (con varias categorías) una variable dicotomizada (que el investigador obligó a ser dicotómica). También mide la relación entre una dicotómica y una dicotomizada. Es un caso particular de la correlación Pearsoniana. No se va a desarrollar porque la mayoría de los autores recomienda utilizar en esta situación la correlación de Pearson.



Correlación entre una Variable Nominal de varias categorías y una Variable Intercalar (u ordinal)

Omega Cuadrado (ω^2)

Este coeficiente de asociación según Weimer (1996, p.624) está indicado en aquellos casos en los cuáles se requiera la asociación en tre una variable nominal y otra (intervalar u ordinal).

Este caso puede presentarse en el ámbito de un ANOVA cuyo valor F halla dado significativo y sabiendo según este valor F que hay una relación entre las dos variables ahora nuestro interés radica en conocer el grado de intensidad de la asociación.

El estadístico omega cuadrado (ω^2) es un estimador común de la fuerza de las asociación entre la variable del tratamiento y la dependiente en un arreglo de ANOVA de un solo criterio de clasificación. Fue derivado por Hays y tiene la siguiente fórmula:

$$\hat{\omega}^2 = \frac{SCTRAT - (k-1)CMERROR}{SCT + CMERROR}$$

Donde:

(ω^2): Omega cuadrado de Hays

SCTRAT: Suma de Cuadrados entre Tratamientos

CMERROR: Cuadrado medio del error

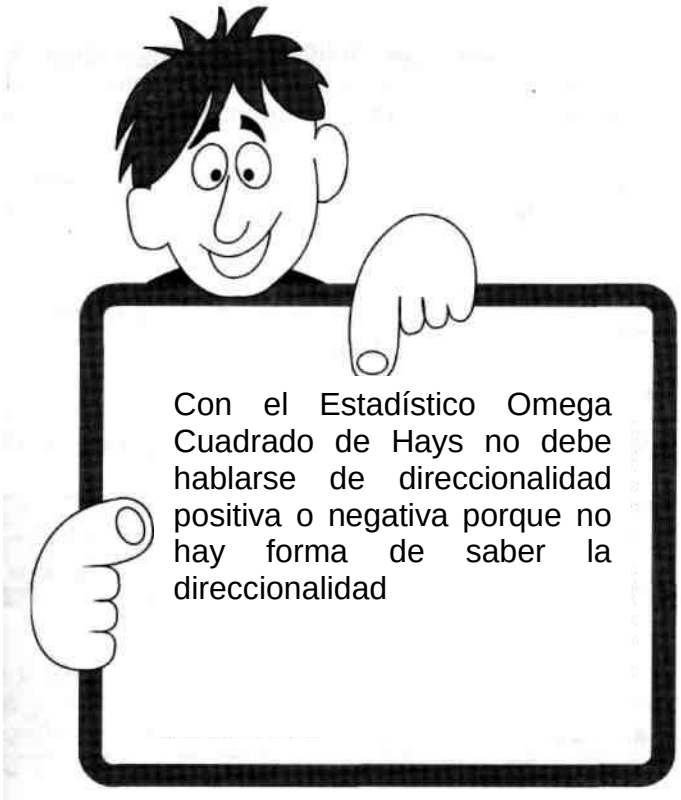
SCT: Suma de cuadrados totales

K: es el número de tratamientos

El estadístico (ω^2) Omega Cuadrado de Hays no está incorporado todavía en algunos software pero la mayoría de ellos provee los insumos necesarios para poder determinarlo en forma indirecta.

Para su interpretación debe utilizarse el siguiente baremo:

Rango (ω^2) de Omega Cuadrado	Intensidad de Relación
0,00 a 0,29	Débil
0,30 a 0,69	Moderada
0,70 a 1,00	Fuerte



Ejemplo de aplicación del Omega Cuadrado de Hays

Se realizó un experimento para determinar:

(a) Si son distintas las medias del número de cirugías de pacientes externos realizadas (por semana) en tres hospitales: General del Sur, Universitario y Coromoto.

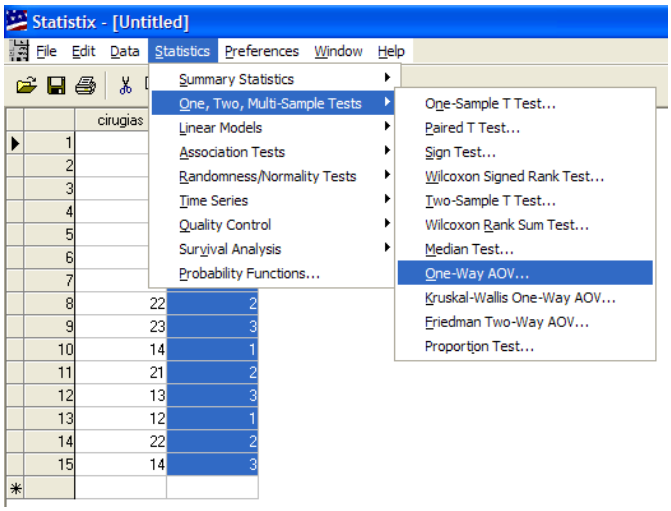
(b) La intensidad de la relación entre el número de cirugías por semana y el tipo de hospital

Hospital General del Sur	Hospital Universitario de Maracaibo	Hospital Coromoto
19	25	25
19	23	23
18	22	23
14	21	13
12	22	14
16,4=Media	22,6=Media	19,6=Media

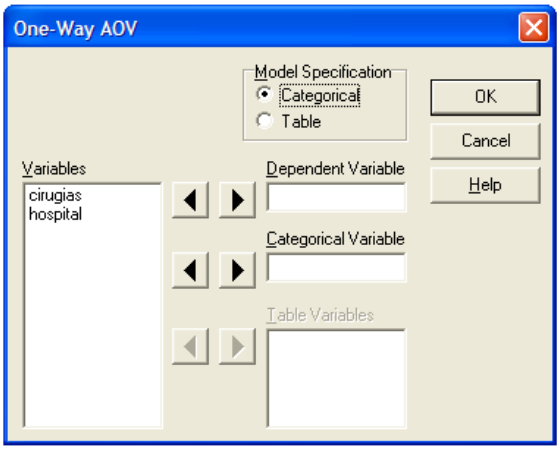
La matriz de datos (en formato categorical) para el Statistix es:

cirugias	hospital
19	1
25	2
25	3
19	1
23	2
23	3
18	1
22	2
23	3
14	1
21	2
13	3
12	1
22	2
14	3

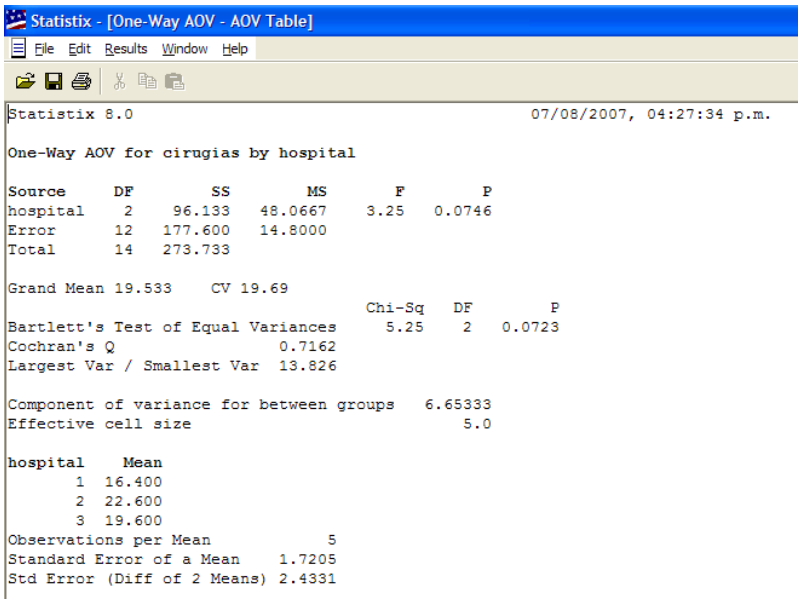
Las opciones del menú son:



La caja de diálogo es:



Aunque la F de Fisher no resultó significativa, se realizará el cálculo del Omega cuadrado como ejemplo didáctico:

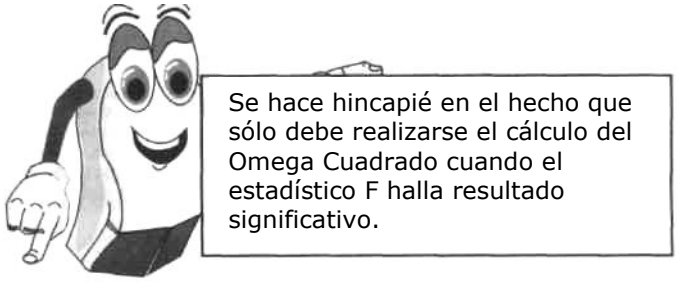


Entonces el Omega Cuadrado es:

$$\hat{\omega}^2 = \frac{SCTRAT - (k-1)CMERROR}{SCT + CMERROR}$$

$$\hat{\omega}^2 = \frac{96{,}13 - (3-1)14{,}8}{273{,}733 + 14{,}8} = \frac{125{,}73}{288{,}533} = 0{,}4357$$

En consecuencia, el 43,57% de la varianza en el número de cirugías puede ser atribuido a la variable hospital.



Correlación Parcial

La correlación parcial se define como la correlación entre dos variables manteniendo las variables intervinientes controladas.

Es muy útil cuando entre las variables no se manifiestas las verdaderas correlaciones a causa de que una tercera variable opaca la correlación entre aquellas dos.

Un ejemplo de este tipo se observa en el efecto mediador que ejercen las calificaciones (variable interviniente) en el efecto que la motivación del alumno tiene sobre la evaluación que hace del profesor; los resultados de las investigaciones educacionales muestran, por ejemplo, que los alumnos con baja motivación hacia el trabajo académico evaluarán desfavorablemente cuando obtienen calificaciones bajas y favorable cuando obtienen altas calificaciones. Por lo tanto, si el interés está dirigido a determinar la relación verdadera o genuina entre motivación y evaluación, será necesario emplear un control estadístico que permita extraer tanto de la

motivación como de la evaluación, el efecto de las calificaciones. Este control se logra calculando el coeficiente de correlación parcial.

La correlación parcial es una estimación de la correlación entre dos variables después de remover de ellas los efectos de otra variables (variable mediadora o interviniente)

Simbólicamente, $r_{y1.2}$ representa la correlación parcial entre Y y X1 después que se ha excluido de ellas el efecto de X2.

La fórmula que se emplea es:

$$r_{Y1.2} = \frac{r_{Y1} - r_{Y2}r_{12}}{\sqrt{(1-r_{Y2}^2)(1-r_{12}^{2)}}}$$

Es decir, en el numerador tenemos:

r_{Y1} :es la correlación simple entre Y y X1

r_{Y2} : es la correlación simple entre Y y X2

r_{12} : es la correlación simple entre X1 y X2

En el denominador figura:

r_{Y2}^2 :es el coeficiente de determinación entre Y y X2

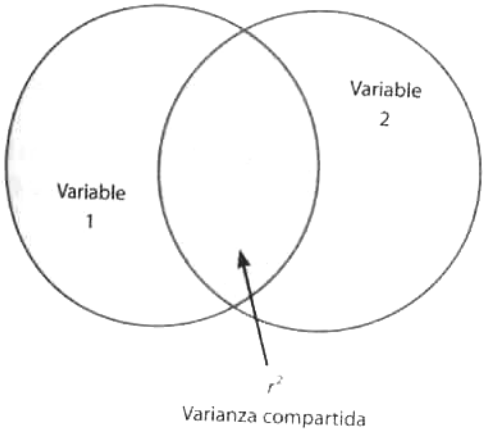
r_{12}^2 : es el coeficiente de determinación entre X1 y X2

A este coeficiente también se le denomina: Coeficiente de Correlación Parcial de Primer Orden, debido a que sólo se controló o parcializó una variable X2. Si se controla por dos variables se denota como:

$r_{Y1.23}$:y se le llama Coeficiente de Correlación Parcial de Segundo Orden y así sucesivamente.

En el modelo de la regresión lineal múltiple, el coeficiente de correlación parcial, se concibe como una relación entre varianzas residuales. En este contexto, también se emplea el coeficiente de correlación parcial al cuadrado.

Para entender mejor el significado del coeficiente de correlación parcial y su correspondiente coeficiente al cuadrado, resulta útil emplear el mismo diagrama que presentan Cohen y Cohen (1983) y que comúnmente es empleado para explicar la noción de varianza compartida.



En el diagrama la varianza de cada variable se representa por un círculo de área igual a la unidad. Las áreas superpuestas de dos círculos representan la relación entre las dos variables, cuantificada por r^2.El área total de Y cubierta por las áreas de X1 y X2 representan la proporción de la varianza de Y que es explicada por dos variables independientes y se representa por R^2. El diagrama muestra, que esta proporción es igual a la suma de las áreas a, b y c.

Las áreas a y b representan aquella porción de Y explicada únicamente por X1 y X2, respectivamente; en tanto que el área c representa la proporción de Y que es explicada simultáneamente por X1 y X2.

El coeficiente de correlación parcial al cuadrado de Y con X1 parcializando o controlando a X2 se expresa así:

$$r^2_{Y1.2} = \frac{a}{a+m} = \frac{R_{Y12} - R^2_{Y.2}}{1 - R^2_{Y2}}$$

Donde $R^2_{Y2} = r^2$, es decir, r_{Y2} es el coeficiente de correlación de orden cero entre Y y X2. Nótese que en la fórmula anterior:

$r^2_{Y1.2} = \frac{a}{a+m} = \frac{R_{Y12} - R^2_{Y.2}}{1 - R^2_{Y2}}$, el $r^2_{Y1.2}$ representa la proporción de la varianza de Y que es estimada o explicada sólo por X1, es decir, el coeficiente de correlación parcial al cuadrado responde a la pregunta: ¿cuánto de la varianza en Y que no es estimada por las otras variables, es estimada por X1?

Ejemplo de Aplicación de Correlación Parcial

A continuación se presenta la información suministrada por 32 supervisores de la empresa MAXY, en la cual se registró por el gerente de recursos humanos, la motivación de logro X1 y la percepción del ambiente de trabajo X2, además de la evaluación del desempeño Y. El gerente pretende verificar la hipótesis siguiente:

"La motivación de logro del supervisor y la percepción que el mismo tiene del ambiente de trabajo, contribuirán de manera importante a explicar la evaluación del desempeño que ellos hacen"

Matriz de Datos

Supervisor	Desempeño Y	Motivación de Logro X1	Percepción del Ambiente X2
01	6	5	4
02	9	7	4
03	4	5	6
04	6	6	9
05	3	3	5
06	8	8	4
07	4	2	5
08	2	2	4
09	10	5	7
10	1	1	3

11	10	8	6
12	6	10	5
13	9	4	7
14	9	3	6
15	6	4	6
16	5	5	7
17	10	9	7
18	7	19	6
19	9	3	5
20	8	4	5
21	10	8	8
22	9	4	6
23	5	10	4
24	2	2	4
25	2	2	5
26	5	5	8
27	10	5	7
28	2	4	6
29	6	4	7
30	8	8	4
31	1	1	2
32	4	5	6

Nota: La escala de valoración fue de 01 a 10

En este ejemplo, las correlaciones simples (orden cero) son:

$r_{Y1} = 0{,}529$

$r_{Y2} = 0{,}447$

$r_{12} = 0{,}247$

El coeficiente de determinación múltiple es:

$R^2_{Y.12} = 0{,}385$

El coeficiente de correlación parcial de Y (desempeño) con motivación de logro (X1) controlando por Percepción del ambiente de trabajo (X2) es:

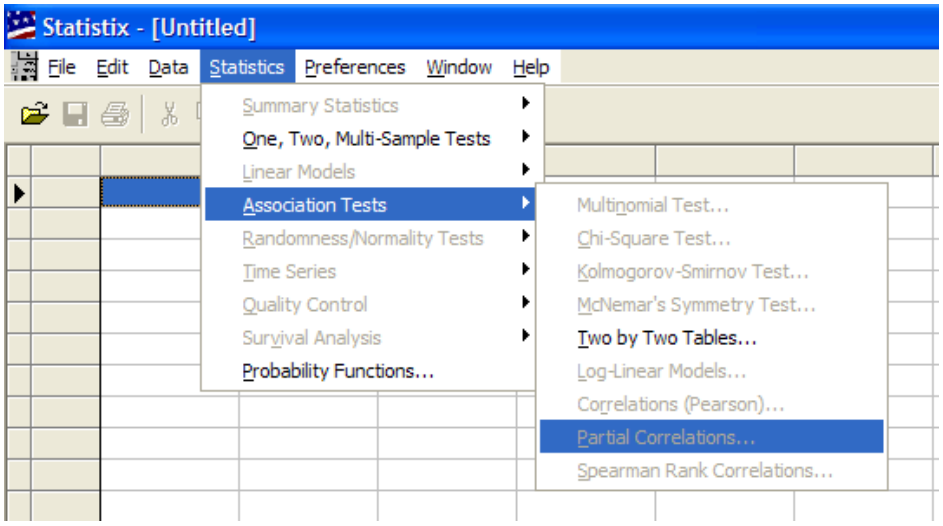
$$r^2_{Y1.2} = \frac{a}{a+m} = \frac{R_{Y12} - R^2_{Y2}}{1 - R^2_{Y2}}$$

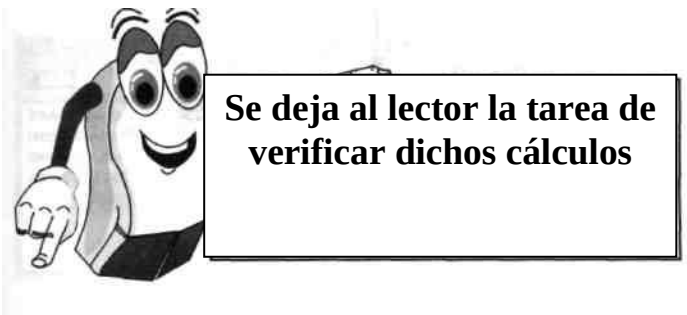
Sustituyendo en la fórmula anterior:

$r^2_{Y1.2} = \frac{0{,}385 - 0{,}199}{1 - 0{,}199} = 0{,}232$, la raíz cuadrada de este valor es la correlación parcial de Y con X1 eliminando la influencia de X2:

$r_{Y1.2} = 0{,}4816$, que representa la verdadera correlación entre Y y X1 después de remover X2 y representa una baja correlación entre evaluación del desempeño y motivación de logro de los supervisores, controlando por el efecto de la percepción del ambiente laboral.

En STATISTIX, la secuencia empleada para obtener estos coeficientes es:

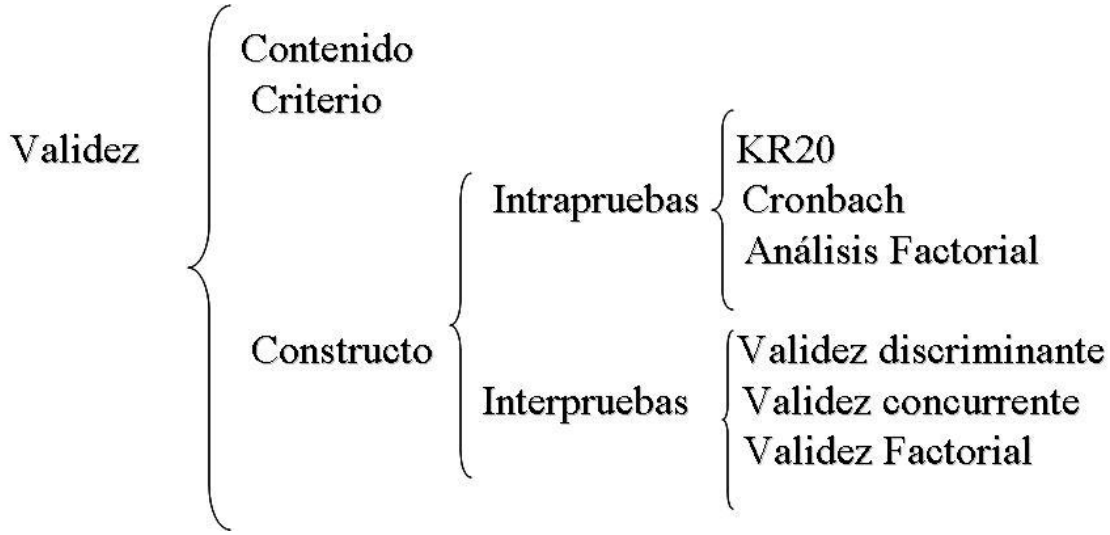


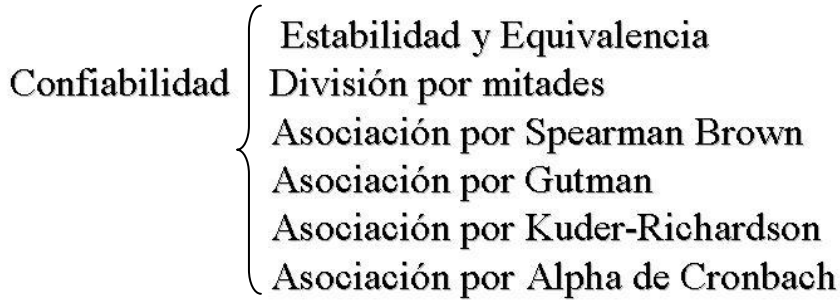


Coeficiente de Confiabilidad Alpha de Cronbach:

Validez
Es la eficacia con que un instrumento mide lo que se pretende medir

Confiabilidad
Es el grado con que se obtienen resultados similares en distintas aplicaciones





Alpha de Cronbach

Sujetos	1	2	3	4	Puntaje Total	$(X - \bar{X})^2$
1	2	3	5	2	12	2.56
2	3	3	1	2	9	1.96
3	5	4	2	3	14	12.96
4	1	1	2	1	5	29.16
5	3	3	2	4	12	2.56
$\bar{X}_i$	2.8	2.8	2.4	2.4	10.4	49.20
S_i^2	1.76	0.96	1.84	1.04	5.6	$S_t^2 = \frac{49.20}{5} = 9.84$

$$Alpha\_Cronbach = \frac{k}{k-1}\left(1 - \frac{\sum S_I^2}{S_T^2}\right) = \frac{4}{4-1}\left(1 - \frac{5.6}{9.84}\right) = 0.57$$

Alpha de Cronbach por matriz de correlación

$$\alpha = \frac{k\bar{p}}{1 + \bar{p}(k-1)}$$

Donde:
K= número de ítemes

$\bar{p}$: Promedio de las correlaciones
(se suman las correlaciones por encima de la diagonal principal y se dividen entre el número de ellas)

Carmines y Zeller (1988) Reliability and validity assessment. Sage Publications, Beverly Hill, California

Coeficiente de Confiabilidad de Hoyt:

El coeficiente de confiabilidad de Hoyt es:

$$r_{tt} = 1 - \frac{S^2_{error}}{S^2_{sujetos}}$$

$$r_{tt} = 1 - \frac{S^2_{error}}{S^2_{sujetos}} = 1 - \frac{1{,}08}{4{,}75} = 0{,}77$$

Donde:
$S^2_{error} = 1{,}08$ (varianza del error)
$S^2_{sujetos} = 4{,}75$ (varianza debida a los sujetos)

De acuerdo a la escala de interpretación dada por Ruiz Bolivar (2002;p.70):

Rangos	Categoría
0,81 a 1,00	Muy alta
0,61 a 0,80	Alta
0,41 a 0,60	Moderada
0,21 a 0,40	Baja
0,01 a 0,20	Muy baja

Se puede concluir, que el instrumento de medición en estudio tiene un coeficiente de confiabilidad alto.

Nota: el coeficiente de Hoyt aparece reseñado en:

Ruiz Bolivar, C. (2002) Instrumentos de Investigación Educativa. Procedimientos para su diseño y validación. Cideg. Lara. Venezuela; pp.68-70

Coeficiente de confiabilidad de Estabilidad y Equivalencia

Sujetos	Pruebas		
	A	B	C
1	17	16	17
2	15	15	15
3	16	14	15
4	12	12	15
5	12	13	13
6	11	10	12
7	11	11	11
8	10	11	13
9	10	10	12
10	10	10	12
11	10	9	10
12	10	10	11
13	9	8	11
14	9	9	11
15	9	9	10
16	8	8	9
17	8	7	10
18	7	7	7
19	6	6	8
20	6	6	8

Coeficiente de Estabilidad

$$r_{AC} = 0,93$$

Coeficiente de Equivalencia

$$r_{AB} = 0,97$$

$$r_{BC} = 0,92$$

Coeficiente de confiabilidad por mitades

Sujetos	Prueba		
	TOTAL	PAR	IMPAR
1	16	8	8
2	14	6	8
3	14	7	7
4	12	6	6
5	13	7	6
6	10	5	5
7	11	5	5
8	11	6	5
9	10	5	5
10	10	5	5
11	9	5	4
12	10	4	6
13	8	4	4
14	9	5	4
15	9	4	5
16	8	4	4
17	7	4	3
18	7	3	4
19	6	3	3
20	6	3	3

Coeficiente de Pares con Impares

$$r_{AC} = 0,77$$

Coeficiente de Confiabilidad
por mitades
(a) Corregido por Spearman-Brown
(Varianza Homogénea)

$$r_{tt} = \frac{2r_{pi}}{1+r_{pi}} = \frac{2(0,77)}{1+0,77} = 0,87$$

(b) Corregido por Gutman
(Varianza Heterogénea)

$$r_{tt} = 2\left(1 - \frac{S_p^2 + S_i^2}{S_T^2}\right) = 2\left(1 - \frac{(1,39^2 + 1,50^2)}{2,75^2}\right) = 0,89$$

Coeficiente de confiabilidad por mitades
¿Cuántos ítemes agregar?

¿Cuántos ítemes?

r_{ttn} : coeficiente de confiabilidad con los ítemes agregados

r_{tt} : coeficiente de confiabilidad antes de agregar ítemes

n: número de ítemes en que se debe aumentar
el instrumento de medición, viene dado por:

$$n = \frac{r_{ttn}(1 - r_{tt})}{r_{tt}(1 - r_{ttn})}$$

Esta fórmula es derivada del método de Spearman-Brown

Coeficiente de confiabilidad por mitades para Diseños Experimentales

Coeficiente de Confiabilidad por Mitades

i. Método de Rulón (Diseño Completamente al Azar)

$$r_{tt} = 1 - \frac{S_d^2}{S_T^2}$$

$$S_d^2 = V_{par} - V_{impar}$$

ii. Método de Hoyt (Diseño de Bloques al Azar)

$$r_{tt} = 1 - \frac{S_e^2}{S_s^2}$$

S_e^2 : Varianza del error

S_s^2 : Varianza entre sujetos

Interpretación del Coeficiente de confiabilidad

Rangos	Magnitud
0,01 a 0,20	Muy baja
0,21 a 0,40	Baja
0,41 a 0,60	Moderada
0,61 a 0,80	Alta
0.81 a 1,00	Muy alta

Ruiz-Bolívar (2002) Elaboración de Instrumentos en Investigación Educativa. CIDEG. Barquisimeto. Lara.

Validez Discriminante de Ítemes

Validez Discriminante de Ítemes

1. Se construye una matriz sujeto-ítemes
2. Se ordenan los puntajes totales de forma ascendente
3. Se calculan los cuartiles Q1, Q2 y Q3
4. Se trabajará con los sujetos situados por encima del Q3 y por debajo del Q1, los demás se eliminan.
5. Se analiza con la prueba t Student cada ítem para ver si existe diferencia significativa entre el grupo bajo y alto.
6. Si resulta, una diferencia significativa entre el grupo alto y bajo para un ítem en cuestión (t sig) se concluye que el referido ítem Discrimina y en dicho caso se deja en el instrumento de medición, caso contrario debe ser eliminado del cuestionario.

Validez Concurrente y Predictiva

- Validez concurrente

Comparar dos situaciones que se miden el mismo momento:
X medición de nuestra situación (mamografia)
Y medición de una situación estándar (biopsia)

- Validez predictiva

X se mide en el momento 1 (coeficiente de inteligencia)
Y se mide en el momento 2 (rendimiento académico)

Cómo Presentar los Cuadros estadísticos

ESTADÍGRAFOS DE LA VARIABLE CLIMA ORGANIZACIONAL

INDICADORES	ITEMS	MEDIA	VARIANZA	DESV. ESTÁNDAR
ESTRUCTURA	1-5	2.41	1.84	1.36
CONFLICTO	6-11	2.97	0.50	0.71
CALIDEZ	12-16	2.75	0.76	0.87
COOPERACIÓN	17-25	3.12	1.88	1.37
PROGRESO	26-32	2.50	3.30	1.81
PROM. INDIVID.	33-35	3.81	0.26	0.51
DIMENSIONES	**ITEMS**	**MEDIA**	**VARIANZA**	**DESV. ESTÁNDAR**
RESTRICTIVO	1-11	2.69	1.29	1.13
AMIGABLE	12-25	2.93	2.28	1.51
ENRIQUECEDOR	26-35	3.15	3.24	1.80
VARIABLE	**ITEMS**	**MEDIA**	**VARIANZA**	**DESV. ESTÁNDAR**
CLIMA ORGANIZ	1-35	2.92	2.44	1.56

Cómo Presentar los Cuadros Estadísticos

ESTADÍGRAFOS DE LA VARIABLE ESTILO DE LIDERAZGO

INDICADORES	ITEMS	MEDIA	VARIANZA	DESV. ESTÁNDAR
SUP. ADMINIST.	1-7	2.55	2.38	1.54
DETERM. ROLES	8-17	2.93	1.93	1.24
PROM. AMISTAD	18-24	3.02	2.45	1.68
TRAB. EQUIPO	25-30	2.98	2.38	1.48
VALOR PRODUCT	31-35	3.62	1.62	1.30
PRO. MET. INDIV.	36-42	3.65	2.91	1.70
ESTILOS	**ITEMS**	**MEDIA**	**VARIANZA**	**DESV. ESTÁNDAR**
FORMAL	1-17	2.74	1.99	1.41
DEMOCRATICO	18-30	3.00	2.54	1.59
APOYO	31-42	3.63	3.41	1.84
VARIABLE	**ITEMS**	**MEDIA**	**VARIANZA**	**DESV. ESTÁNDAR**
ESTILO LIDERAZ	1-42	3.12	2.36	1.62

Uso de la Razón de Acuerdos

Comparación de las actitudes de hombres y mujeres frente a políticas de control de cambio.

	Hombres	*Mujeres*	*Total*
Acuerdo	*180*	*100*	*280*
Desacuerdo	*70*	*140*	*210*
Razón acuerdo/desacuerdo	*2.57*	*0.71*	*1.33*

Capítulo IX

Análisis de Regresión no lineal

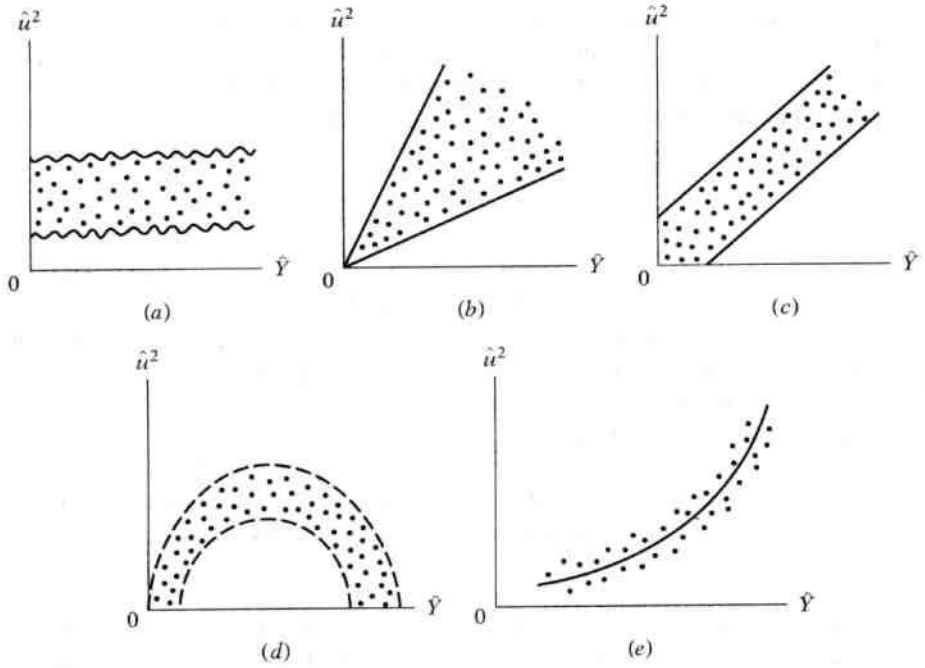
En algunas situaciones es evidente que un modelo lineal en todas las variables independientes es inadecuado. Por ejemplo, un modelo de regresión que predice Y, la puntuación de las preferencias en una prueba de sabor de una bebida de lima, como función lineal de X1 (concentración de lima) y X2 (dulzura) es muy dudosa. En otros casos, los diagramas de dispersión de los datos revelan la ausencia de linealidad. Particularmente, en la regresión lineal, una gráfica ordinaria de Y contra X es adecuada.

En la regresión múltiple, el efecto de las variables entre sí puede oscurecer la ausencia de linealidad. Por esta razón, una estrategia estándar es ajustar un modelo de primer orden y después representar gráficamente los residuos de este modelo con cada variable independiente. Si un modelo más apropiado contiene, por ejemplo, un término en segundo grado X_1^2, entonces la gráfica de los residuos ($\varepsilon = Y_i - \hat{Y}$) contra X1 muestra un patrón no lineal. Como el uso de un modelo de primer orden elimina el efecto lineal de las otras variables independientes, las no lineales a menudo se muestran con mayor claridad en estos diagramas residuales.

Las interacciones entre las variables son más difíciles de descubrir en los diagramas de dispersión. Si X1 y X2 interactúan al determinar Y, hay tres variables involucradas; desafortunadamente, los diagramas tridimensionales son más difíciles de trazar e interpretar. Quizá el sentido común sea la mejor manera de determinar si hay interacciones presentes.

Para el caso especial en que una de las variables independientes es una variable cualitativa representada por una o más variables ficticias (dummy), la interacción se puede descubrir trazando gráficas de los residuos (tomados de un modelo de primer orden) contra otras variables independientes. Se deberían trazar gráficas por separado para las observaciones en cada categoría de la variable cualitativa. El modelo de primer orden sin interacciones implica que estos diagramas deben ser paralelos. Si los diagramas de los residuos por separado no son más o menos paralelos, se deberían considerar la posible presencia de algún tipo de interacción.

La representación gráfica de estos residuales con respecto al valor estimado de Y, pueden producir los siguientes patrones:

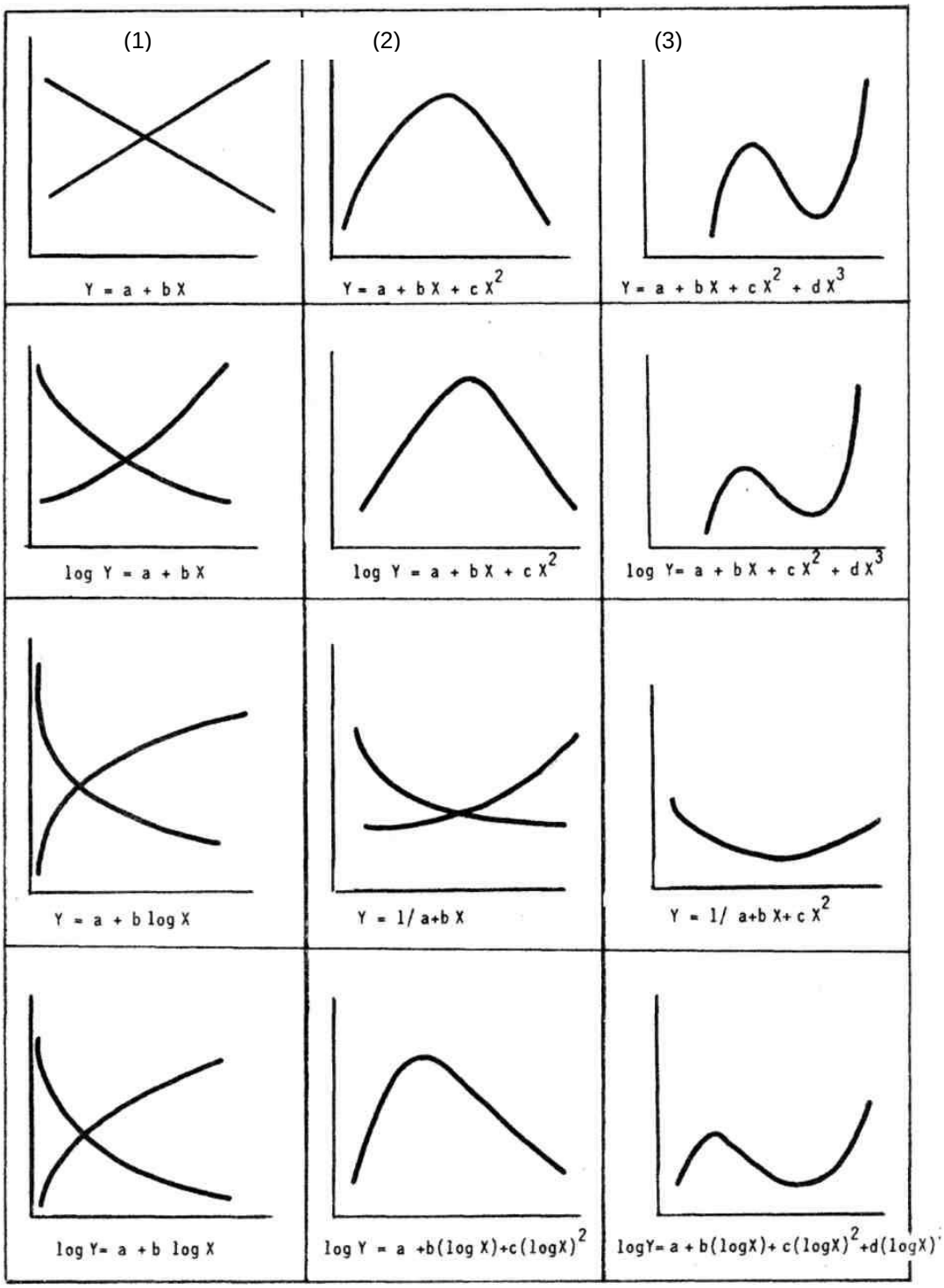


La idea de estos gráficos es averiguar si el valor medio estimado de Y está relacionado sistemáticamente con el residual al cuadrado.

En (a) se ve que no hay un patrón sistemático entre las dos variables, lo cual sugiere que posiblemente no hay heteroscedasticidad en los datos. Sin embargo las gráficas de (b) a (e) muestran patrones

definidos. Por ejemplo, la figura (c) sugiere una relación lineal, mientras que (d) y (e) indican una relación cuadrática. Utilizando estas gráficas, es posible obtener curvas de mejor ajuste u orientarnos acerca del tipo de transfomación que debe aplicarse a los datos a fin de mejorar el ajuste. Pueden también hacerse representaciones de los residuales versus el valor de X.

Los modelos de regresión no lineal son muy variados. Algunos de estos modelos se dan a continuación:


(1)
(2)
(3)
Y = a + b X
Y = a + b X + c X^2
Y = a + b X + c X^2 + d X^3
log Y = a + b X
log Y = a + b X + c X^2
log Y = a + b X + c X^2 + d X^3
Y = a + b log X
Y = 1/ a+b X
Y = 1/ a+b X+ c X^2
log Y = a + b log X
log Y = a +b(log X)+c(logX)^2
logY = a + b(logX)+ c(logX)^2+d(logX)^3

En la columna (1) aparecen de arriba abajo el modelo lineal, el semilogarítmico inverso en Y, el modelo inverso logarítmico en X, doble logarítmico. En la columna (2) aparecen el modelo cuadrático, el modelo cuadrático inverso logarítmico en Y, el modelo recíproco, el modelo de parábola logarítmica y en la tercera columna aparecen modelos polinomiales.

Existen una gama muy amplia de modelos. Pero se pueden clasificar en dos grupos los intrínsecamente lineales, es decir, se pueden transformar a la forma lineal y los no lineales propiamente dichos que no se pueden transformar a la forma lineal. Los métodos de estimación cambian en ambos casos en los intrínsecamente lineales puede aplicarse la transformación y luego aplicar el método de los mínimos cuadrado. En los no lineales propiamente dichos se aplica el método de estimación de máxima verosimilitud.

Entre los modelos intrínsecamente lineales más usuales en la práctica tenemos:

(a) **El modelo lineal** útil en los casos de ajuste no lineal porque sirve como patrón de comparación.

Ecuación del modelo $\hat{Y} = a + bX$

(b) **El modelo recíproco** (también llamado hipérbola) donde una de las variable va aumentado y la otra va disminuyendo.

Ecuación del modelo $\hat{Y} = \frac{1}{a + bX}$

(c) El modelo gamma que es muy utilizado en ajustes de curvas de producción lechera.

Ecuación del modelo $\hat{Y} = ab^X X^c$ su transformación a la forma lineal es:

logY=log(a)+x(logb) + c(logX)

(d) El modelo potencial muy utilizado en ajusten de precio-demanda.

Ecuación del modelo $\hat{Y} = aX^b$ su transformación a la forma lineal es:

Log Y=log(a) + b(logX)

(e) El modelo exponencial muy utilizado en ajustes de crecimiento poblacionales.

Ecuación del modelo $\hat{Y} = ab^X$ su transformación a la forma lineal es:

logY = log(a)+X(logb)

Entre los modelos no lineales propiamente dichos encontramos:

(a) El modelo logístico para estudiar el crecimiento de poblaciones.

Ecuación del modelo $\hat{Y} = \frac{1}{k + ab^X}$ o $\frac{1}{Y} = ab^X + k$

(b) El modelo de parábola logarítmica, cuya ecuación del modelo es:

$$\hat{Y} = a + b(\log X) + c(\log X)^2$$

(c) El modelo de Gompertz también usado para el estudio de crecimientos poblacionales.

Ecuación del modelo: $\hat{Y} = pq^{b^x}$ o $\log Y = \log(p) + b^x \log(q)$

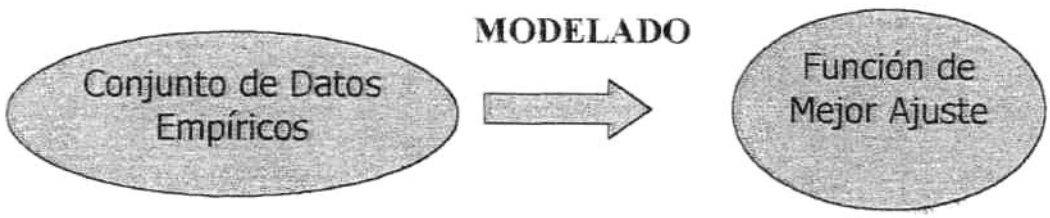
Modelaje, Modelado o Modelización

Es la construcción de un modelo que se sabe a ciencia cierta, la respuesta que produce, luego se aplica a una situación donde dicha respuesta es desconocida para conocer su comportamiento, y, si el modelo resulta ser de "buen ajuste en lo sucesivo lo aplicaremos para propósitos de predicción.

Modelado Matemático-Estadístico: es una función que describe un proceso o fenómeno.

El tipo o función queda determinado por:(a) Gráfica de la función;

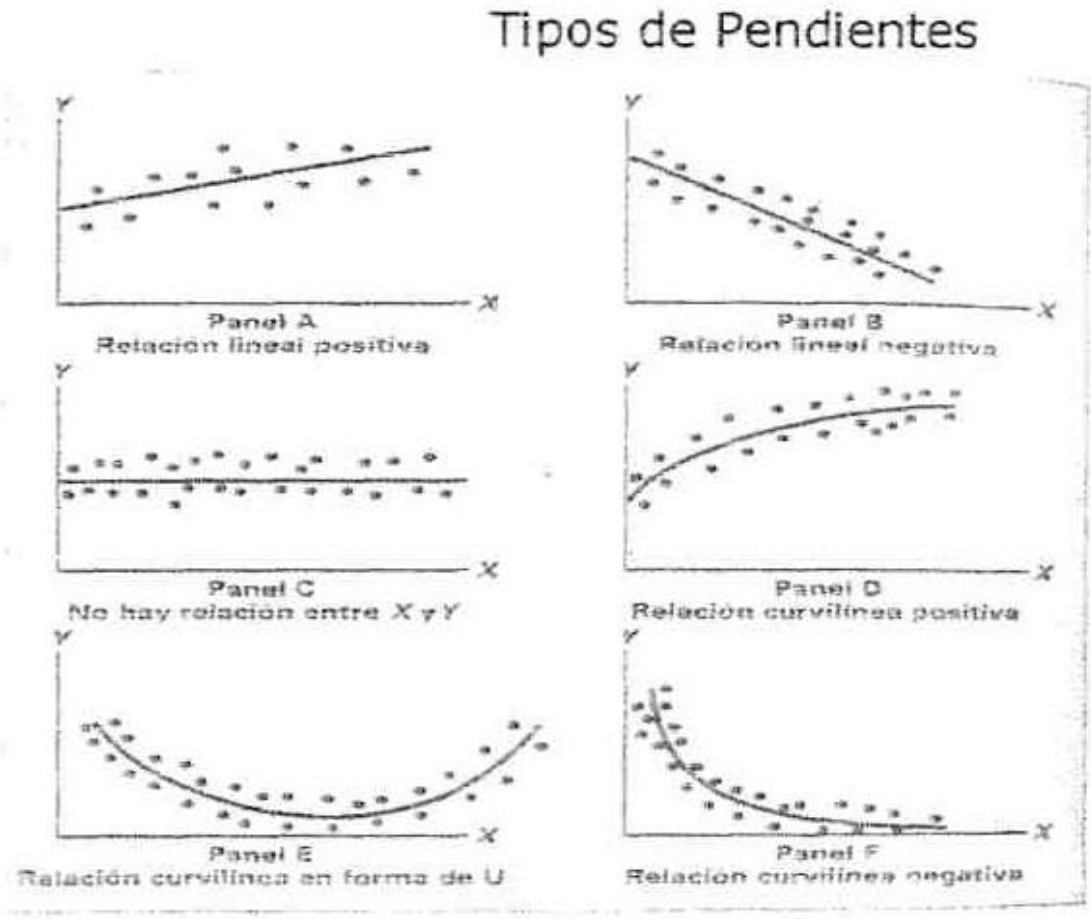
(b) Algún Principio subyacente a los datos(c)Mejor Criterio de ajuste.



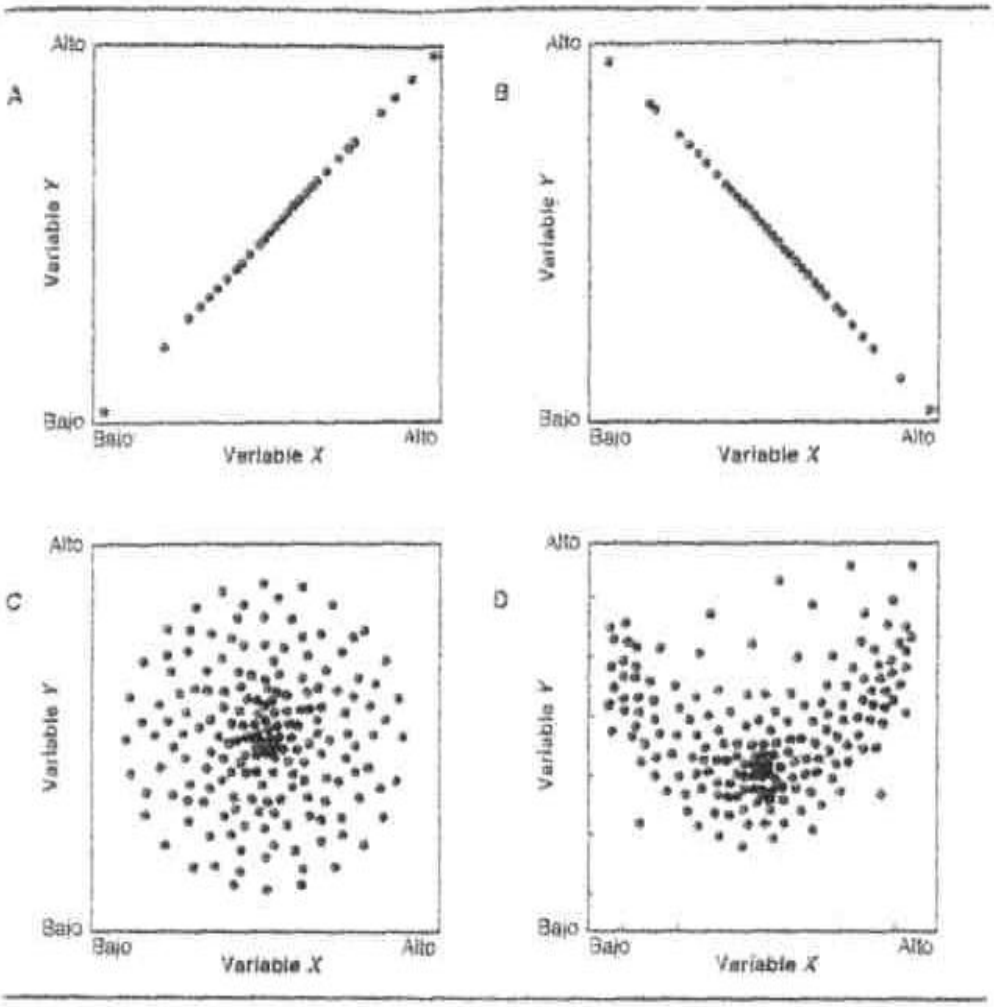
Buen ajuste

Es el grado de semejanza entre las respuestas reales (Yi) y las obtenidas con el modelo

Tipos de pendientes

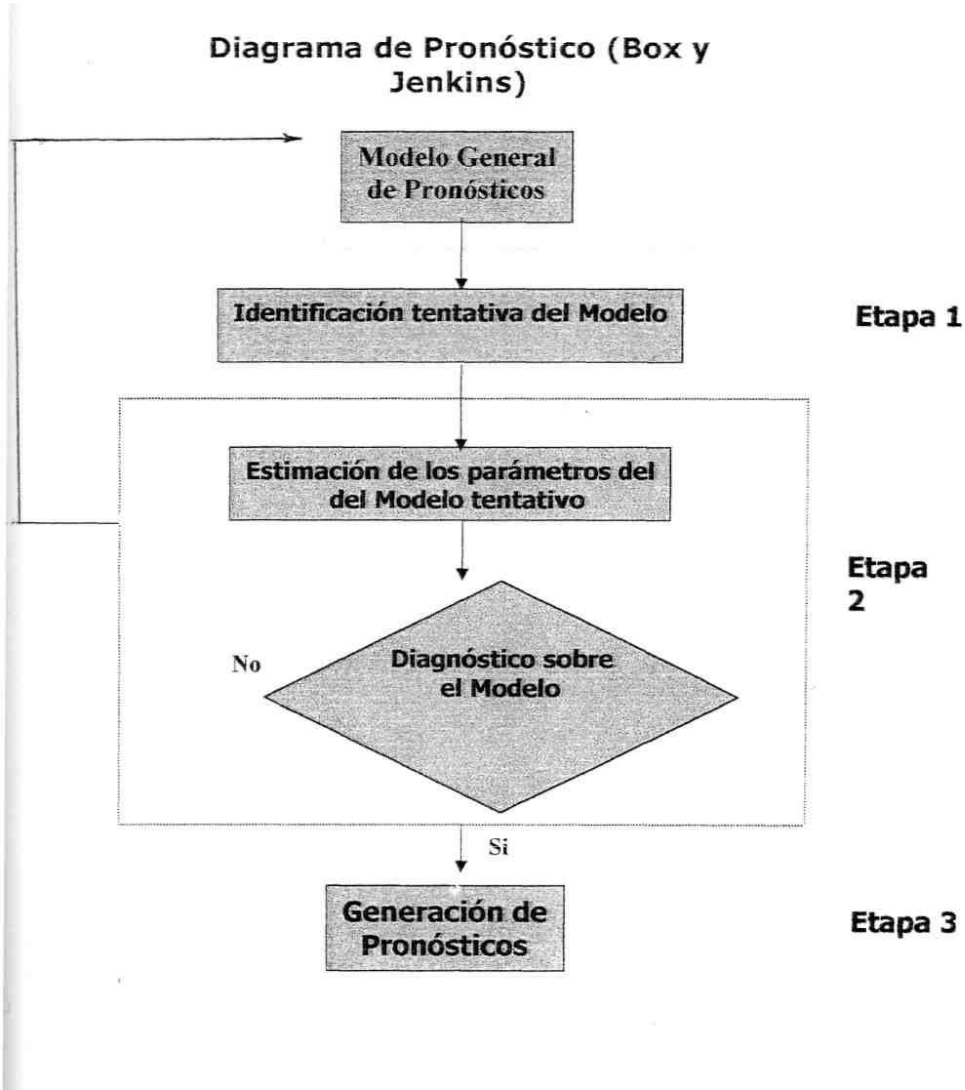


Tipos de Diagramas de Dispersión


A
Alto
Variable Y
Bajo
Bajo
Alto
Variable X
B
Alto
Variable Y
Bajo
Bajo
Alto
Variable X
C
Alto
Variable Y
Bajo
Bajo
Alto
Variable X
D
Alto
Variable Y
Bajo
Bajo
Alto
Variable X

Definición de Mejor Modelo:

- Más fiable (o confiable)
- Menor error estándar en los estimadores
- Más sencillo (Principio de parsimonia): Menos parámetro
- Menos suposiciones para construirlo



Pautas para usar la ecuación de regresión lineal

(1) Si no hay una correlación lineal significativa, no use !a ecuación de regresión para hacer predicciones.

(2) No use la ecuación de regresión para hacer predicciones fuera del rango de valores muestreados.

(3) Una ecuación de regresión basada en datos viejos (obsoletos) no necesariamente sigue siendo válida en el presente.

(4) No haga predicciones acerca de una población distintas de la población de la cual se extrajo la muestra de datos.

Ejemplo de Aplicación

Mediante un ejemplo relacionado con el crecimiento (longitud:Y) de una larva de un parásito por días de eclosión (X) se ajustarán una serie de modelos y se escogerá el que sea de "mejor ajuste" en base al criterio del R-cuadrado.

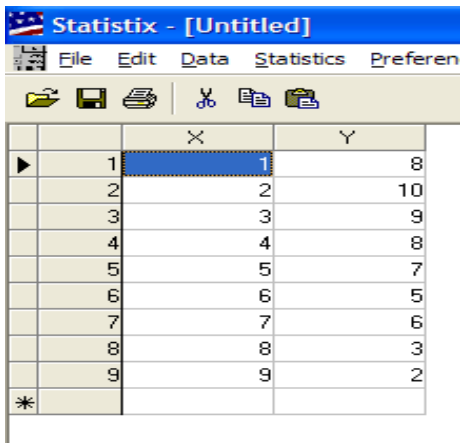
X (días)	1	2	3	4	5	6	7	8	9
Y (longitud)	8	10	9	8	7	5	6	3	2

Se pide ajustar los modelos siguientes:

(a) Lineal
(b) Exponencial
(c) Potencial
(d) Recíproco
(e) Inverso
(f) Gamma
(g) Polinomial trigonométrico
(h) Parábola
(i) Parábola de raíz
(j) Logarítmico en X
(k) Logarítmico en Y
(l) Cúbico

Solución

(a) Ajuste del modelo lineal: comenzamos por vaciar los datos en el paquete STATISTIX.

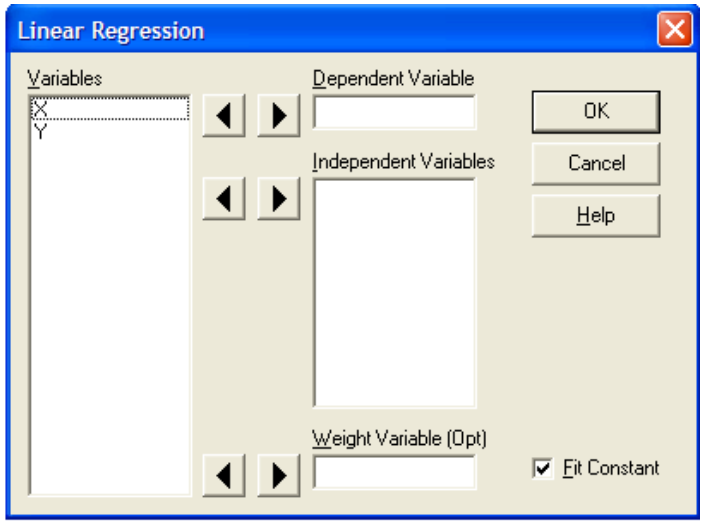


		X	Y
▶	1	1	8
	2	2	10
	3	3	9
	4	4	8
	5	5	7
	6	6	5
	7	7	6
	8	8	3
	9	9	2
*			

Luego se sigue la secuencia de instrucciones

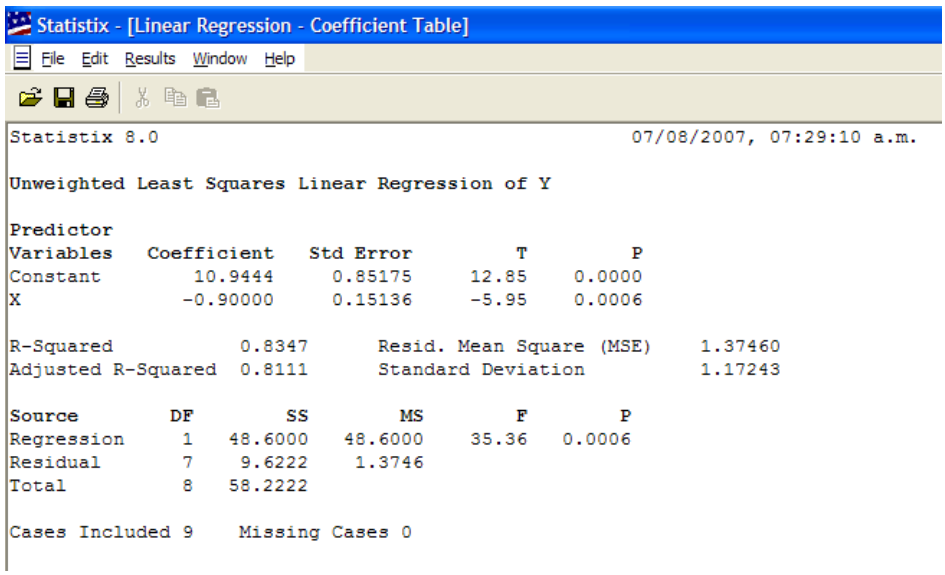
Statistix >linear models>linear regression

Seguidamente aparece el siguiente menu de dialogo.



Declaramos Y como variable dependiente y X como la independiente.

Al presionar okay se produce la salida siguiente.



El modelo lineal es:

Y=10,9444 - 0,90000X con un $R^2 = 0{,}8347$

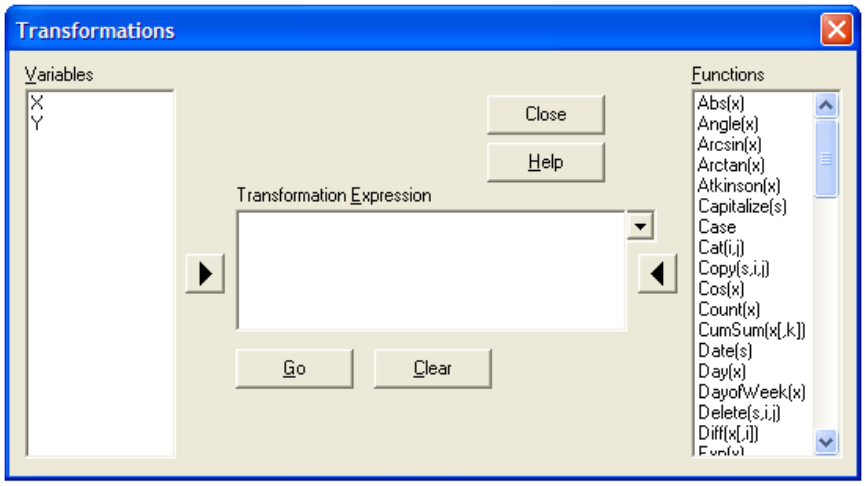
Podemos reescribir la ecuación así:

Longitud=10,9444 - 0,9 Días con un $R^2 = 0{,}8347$

(b)Ajuste del modelo exponencial. Como este modelo tiene la siguiente expresión:

$Y = ab^x$ para expresarlo en la forma lineal es:

logY=log(a)+X(logb), es decir, para introducir la data en STATISTIX debemos crear la variable LOG(Y). En este momento acudimos a la opción TRANSFORMATION del menú DATA y aparece la siguiente caja de dialogo:



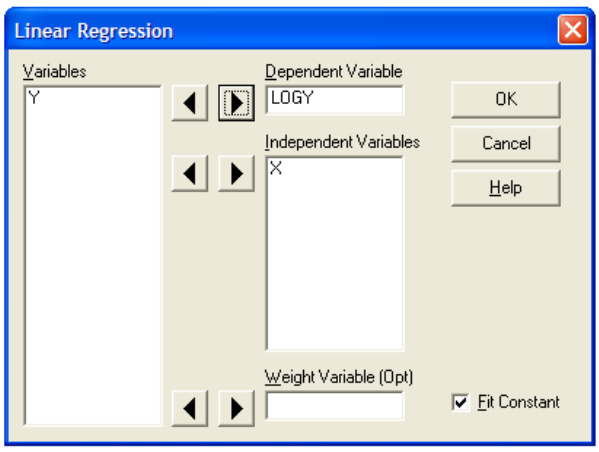
Y en el interior de la caja que dice Transformation Expresión Escribimos:

LOGY=log(Y)

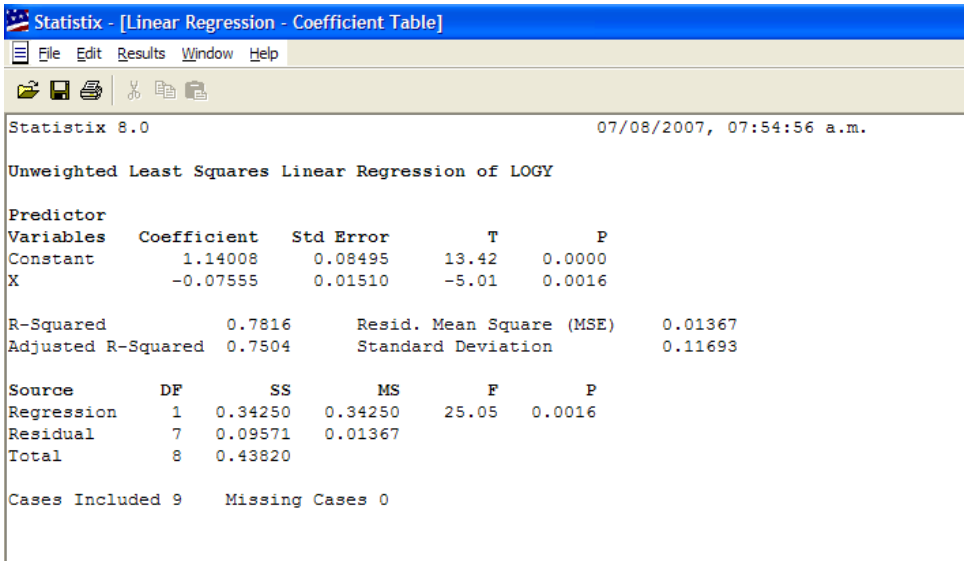
Es decir, estamos obteniendo el logaritmo de Y y se lo estamos asignando a una variable llamada LOGY. Procedemos a presionar GO y se crea esta variable; a continuación volvemos a invocar el procedimiento para la regresión:

STATISTICS>LINEAR MODELS>LINEAR REGRESSION

A continuación aparece y pasamos con el cursor LOGY como variable dependiente y X como independiente así:



Al presionar OKEY aparece la salida:



Luego la ecuación del modelo es:

LOGY=1,14008 – 0,07555 X con un $R^2 = 0{,}7816$

Para no hacer tan extenso el procedimiento, por demás repetitivo, para el modelo potencial (se crea la variable LOGX y LOGY), para desarrollar el modelo recíproco se crea la variable inversa de Y (invY=1/y), para desarrollar el modelo inverso creamos la variable inverso de X (invX=1/X), para desarrollar el modelo gamma se crean las variables: LOGY y LOGX, para el modelo polinomial trigonométrico se crean dos variables SENOX y COSX, así:

SENOX=sin(2*PI*X/24) y COSX=cos(2*PI*X/24), para la parábola se crea la variable X cuadrado X2=X*X, para desarrollar el modelo parábola de raíz se crea la variable RAIZX=SQRT(X), para desarrollar el modelo cúbico se crea la variable X cubo, así X3=X^3 ó X3=X*X*X.

El desarrollo de estos modelos de regresión, aparecen resumido en el cuadro que sigue mostrando cada modelo y su respectivo R-cuadrado:

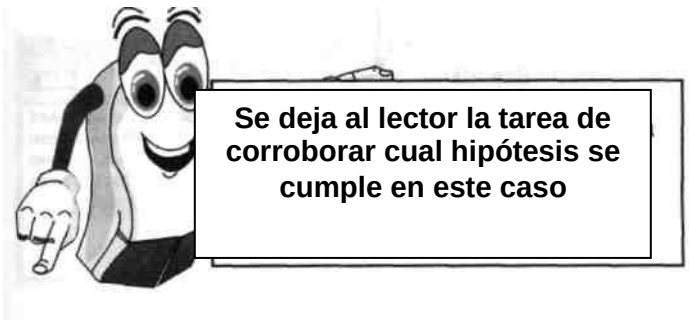
Modelo	R-Cuadrado
1.Lineal	0,8608
2.Cuadrático	0,8632
3.Logarítmico en X	0,6544
4.Cúbico	0,9643
5.Inverso	0,2002
6.Potencial	0,8560
7. Exponencial	0,8590
8.Recíproco	0,8613
9.Gamma	0,9256
10.Polinomial trigonométrico	0,9238

Se aprecia que entre todos los modelos el de mejor ajuste es el modelo de regresión cúbico por poseer el mayor R-cuadrado.

Restaría verificar si el modelo cúbico es bueno para predecir o estimar comparado los valores de Y observados con los de Y esperados y esto puede hacerse utilizando el Ji-cuadrado. En este sentido, se plantea el sistema hipotético:

Ho: Los valores de Y observados concuerdan con los esperados, y por tanto, el modelo es bueno para predecir.

H1: Los valores de Y observados no concuerdan con los esperados, y por tanto, el modelo es malo para predecir.

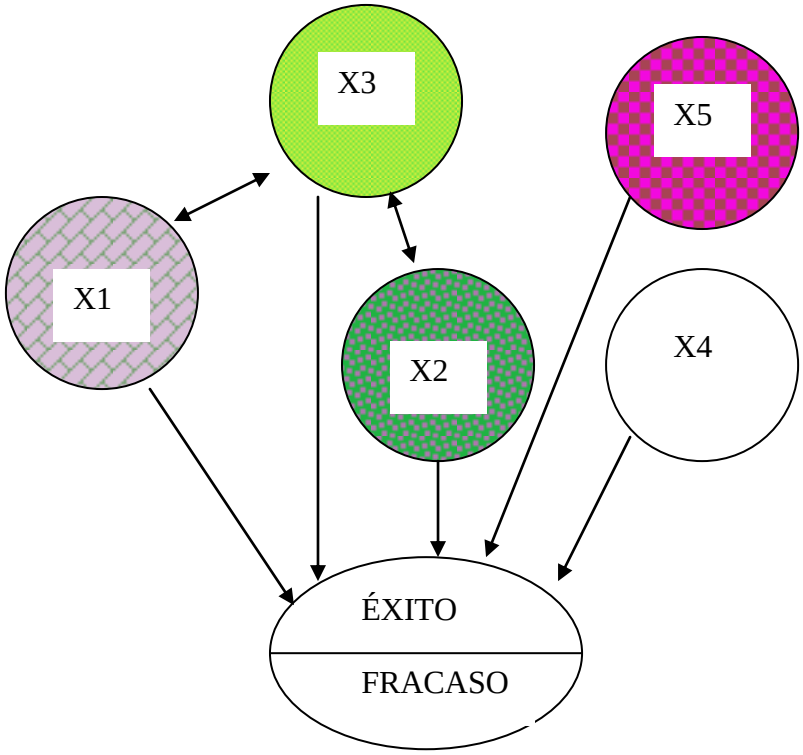


Capítulo X

Regresión Logística

Conceptos y definiciones

Es un modelo clásico de regresión lineal simple o múltiple, pero donde la variable dependiente es binaria o dicotómica.



Es decir, adopta sólo dos valores posibles: éxito y fracaso, positivo y negativo, muerto y vivo, buen y mal desempeño, parasitado o no, aprobado o no aprobado.

La regresión logística es un tipo especial de regresión que se utiliza para explicar y predecir una variable categórica binaria (dos grupos) en función de varias variables independientes que a su vez pueden ser cuantitativas o cualitativas.

Permite modelizar la probabilidad de que ocurra un evento dado una serie de variables independientes.

Razones para utilizar la Regresión Logística

(1) La razón Odds Ratio es una variable discreta (dicotómica) cuyo comportamiento sigue una distribución binomial, invalidando el supuesto básico de normalidad.

(2) La Función de Relación es una regresión intrínsecamente no lineal.

(3) La varianza de una variable dicotómica no es constante, al cambiar los valores de las Xi los puntos de Y se abren en un abanico que refleja la heterocedasticidad.

Aplicaciones de la Curva Logística

- En Economía
  * Podemos querer distinguir entre riesgo de crédito alto y bajo.
  * Empresa rentable o no rentable.
  * Empresa bajo riesgo financiero o no.
  * Éxito de Ventas frente a fracaso en ventas.
  * Compradores (consumidores) frente a no compradores.

- En Veterinaria:
  V. Dependiente (alcanza, no alcanza el peso al destete)
  V. Independientes: Raza, Peso al Nacer, Ganancia de peso, Indice de Quetelet.

Expresiones de la Regresión Logística

El valor teórico recibe diferentes nombres (Sinónimos) en la Literatura Científica:

(1) Odds Ratio
(2) Razón de ventajas
(3) Razón de oportunidades
(4) Razón de desigualdades
(5) Razón de momios
(6) Transformación logística
(7) Razón de verosimilitud
(8) Cociente de posibilidades
(9) Oportunidad Relativa

Odds Ratio:

$$Odds\_Ratio = \frac{P}{Q}$$

$$Odds\_Ratio = \frac{probabilidad\_éxito}{1 - probabilidad\_éxito}$$

Este cociente expresa que si P=0,50 entonces el cociente vale uno:

$$Odds\_Ratio = \frac{0,50}{1-0,50} = 1$$

O un éxito es a un fracaso (1 a 1)

Si P=0,75 entonces el cociente es:

$$Odds\_Ratio = \frac{0,75}{1-0,75} = 3$$

O tres éxitos por un fracaso (3 a 1)

El modelo logístico se basa en el logaritmo natural de este cociente

Regresión logística

$$Odds(E) = e^{\alpha+\beta}$$

$$Odds(\overline{E}) = e^{\alpha}$$

$$OR = \frac{e^{\alpha+\beta}}{e^{\alpha}} = e^{\beta}$$

$$\ln(OR) = \beta \Rightarrow (-\infty < OR < +\infty)$$

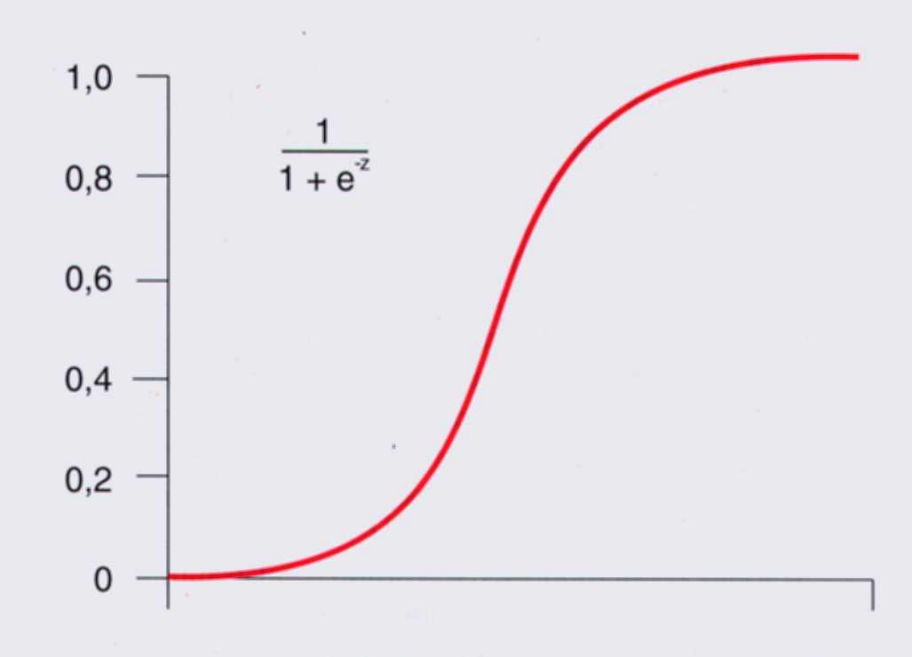
Modelo sin interacción

$$\ln\left(\frac{p}{1-p}\right) = bo + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 X_1 + b_2 X_2 + b_3 (X_1 * X_2) + b_4 X_2^2$$

Curva Logística

Si Z=0 entonces Y= 0.5
Si Z tiende a + infinito entonces Y= 1
Si Z tiende a - infinito entonces Y= 0



Interpretación de los coeficientes estimados Betas:

- Un coeficiente beta positivo aumenta la Odds Ratio (OR)(es decir, la probabilidad de ocurrencia del suceso aumenta).
- Un beta negativo disminuye la OR

Ejemplo de aplicación

Los datos que se dan a continuación pertenecen a 79 niños afectados de enfermedad hemolítica neonatal, de los cuales 63 sobrevivieron y 16 murieron. En cada niño se registró la concentración de hemoglobina en el cordón umbilical, X (medida en gramos por cien mililitros) y la concentración de bilirrubina, Y

((mg/100ml). Queremos predecir, mediante estos dos valores, si un niño determinado tiene más probabilidad de sobrevivir o de morir.

Grupo (1=sobrevivientes; 2=fallecidos)

X=hemoglobina

Y=bilirrubina

Grupo	Hb	BILI
1	18.7	2.2
1	17.8	2.7
1	17.8	2.5
1	17.6	4.1
1	17.6	3.2
1	17.6	1.0
1	17.5	1.6
1	17.4	1.8
1	17.4	2.4
1	17.0	0.4
1	17.0	1.6
1	16.6	3.6

Grupo	Hb	BILI
1	15.4	2.2
1	15.3	2.0
1	15.3	2.0
1	15.1	3.2
1	14.8	1.8
1	14.7	3.7
1	14.7	3.0
1	14.6	5.0
1	14.3	3.8
1	14.3	4.2
1	14.3	3.3

Grupo	Hb	BILI
1	14.3	3.3
1	14.1	3.7
1	14.0	5.8
1	13.9	2.9
1	13.8	3.7
1	13.6	2.3
1	13.5	2.1
1	13.4	2.3
1	13.3	1.8
1	12.5	4.5
1	12.3	5.0

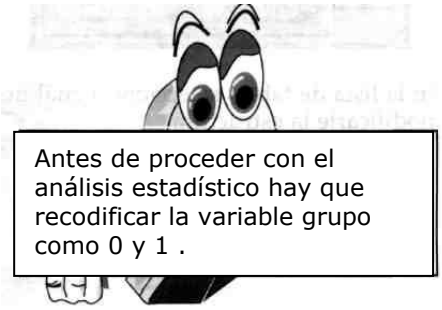
Grupo	Hb	BILI
1	12.2	3.5
1	12.2	2.4
1	12.0	2.8
1	12.0	3.5
1	11.8	2.3
1	11.8	4.5
1	11.6	3.7
1	10.9	3.5
1	10.9	4.1
1	10.9	1.5
1	10.8	3.3

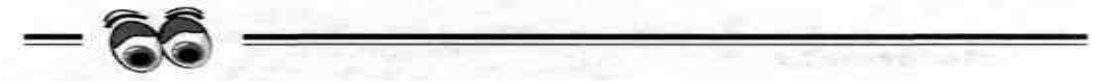
1	16.3	4.1
1	16.1	2.0
1	16.0	2.6
1	16.0	0.8
1	15.8	3.7
1	15.8	3.0
1	15.8	1.7
1	15.6	1.4
1	15.6	2.0
1	15.6	1.6
1	15.4	4.1

1	10.6	3.4
1	10.5	6.3
1	10.2	3.3
1	9.9	4.0
1	9.8	4.2
1	9.7	4.9
1	8.7	5.5
1	7.4	3.0
1	5.7	4.6
2	15.8	1.8
2	12.3	5.6

Grupo	Hb	BILI
2	9.5	3.6
2	9.4	3.8
2	9.2	5.6
2	8.8	5.6
2	7.6	4.7
2	7.4	6.8
2	7.1	5.6
2	6.7	5.9
2	5.7	6.2
2	5.5	4.8
2	5.3	4.8

Grupo	Hb	BILI
2	5.3	2.8
2	5.1	5.8
2	3.4	3.9





Secuencia de Instrucciones en el STATISTIX
STATISTICS>LINEAR MODELS>LOGISTIC REGRESSION

Student Edition of Statistix 8.0 ANEMIA

Unweighted Logistic Regression of CATEGO

Predictor Variables	Coefficient	Std Error	Wald Coef/SE	P
Constant	-2.36928	2.47984	-0.96	0.3394
BILI	-0.49530	0.35580	-1.39	0.1639
Hb	0.53762	0.15534	3.46	0.0005

Deviance	40.06
P-Value	0.9999
Degrees of Freedom	78

Convergence criterion of 0.01 met after 5 iterations

Cases Included 81 Missing Cases 0

Ecuación de regresión logistica

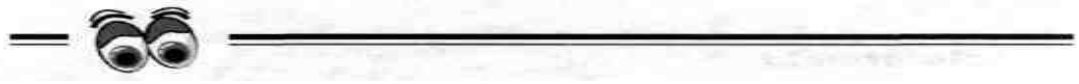
$$sobrevivencia = \beta_0 + \beta_1 Bili + \beta_2 Hb$$

$$Sobrevivencia = -2{,}30 - 0{,}495 Bili + 0{,}53 Hb$$

Del cuadro de ANOVA se concluye que la hemoglobina es un factor pronóstico importante; la bilirrubina no.

Bondad de ajuste del modelo

La bondad de ajuste del modelo se prueba con el estadístico de Desviación. El estadístico de Desviación sigue una distribución ji-cuadrado con n-k-1 grados de libertad. Dicho estadístico compara el modelo actual con el modelo saturado.



Modelo Saturado:

Es aquel que tiene tantos parámetros como puntos de datos

Se plantea un sistema hipotético constituido por hipótesis nula y alterna:

Ho: el modelo es de buen ajuste a los datos

H1: el modelo no es de buen ajuste

Utilizando un nivel alpha de significación, la regla de decisión es:

Rechace Ho si la Desviación es significativa (esto es, está asociada a un valor P<0,05). El cuadro de ANOVA dado muestra que:

Deviance	40.06
P-Value	0.9999
Degrees of Freedom	78

Lo cual revela, que no se rechaza Ho. En consecuencia, el modelo es bueno.

Contribución relativa de cada variable independiente

A continuación, tras haber probado la bondad del modelo nos queda por evaluar el grado de contribución de cada una de las variables independientes al modelo.

El estadístico de prueba está basado en el cociente entre el coeficiente de regresión y el error estándar del coeficiente de regresión. En la regresión logística, a este cociente se le llama: **Estadístico de Wald** que sigue una distribución normal.

	Wald=Coef/SE	P
Constant	-0.96	0.3394
BILI	-1.39	0.1639
Hb	3.46	0.0005

Se aprecia que la bilirrubina no es significativa, es decir,no hace una contribución significativa al modelo; la hemoglobina sí.

Esto nos debe llevar a reflexionar sobre si la Bilirrubina debe continuar en el modelo pues su aporte no es significativo. En otras palabras, la bilirrubina debe ser excluida del modelo.

La Regresión Logística como función discriminante

A partir de las medias de X e Y; de la matriz de varianzas-covarianzas se puede obtener una función discriminante lineal (parecida a la función discriminante de Fisher, que es una combinación lineal de las variables involucradas).

Las medias y sus diferencias son:

		$\bar{X}$	$\bar{Y}$
Sobrevivientes		13.897	3.090
Fallecidos		7.756	4.831
Diferencia	(S – F)	6.141	-1.741
Media	(S+F)/2	10.827	3.961

Sustituyendo en la función discriminante los promedios, obtenemos:
Z = 0.6541 X - 0.3978 Y

Z = (0.6541)(10.827)+(-0.3978)(3.961) = 5.506 (punto de corte)

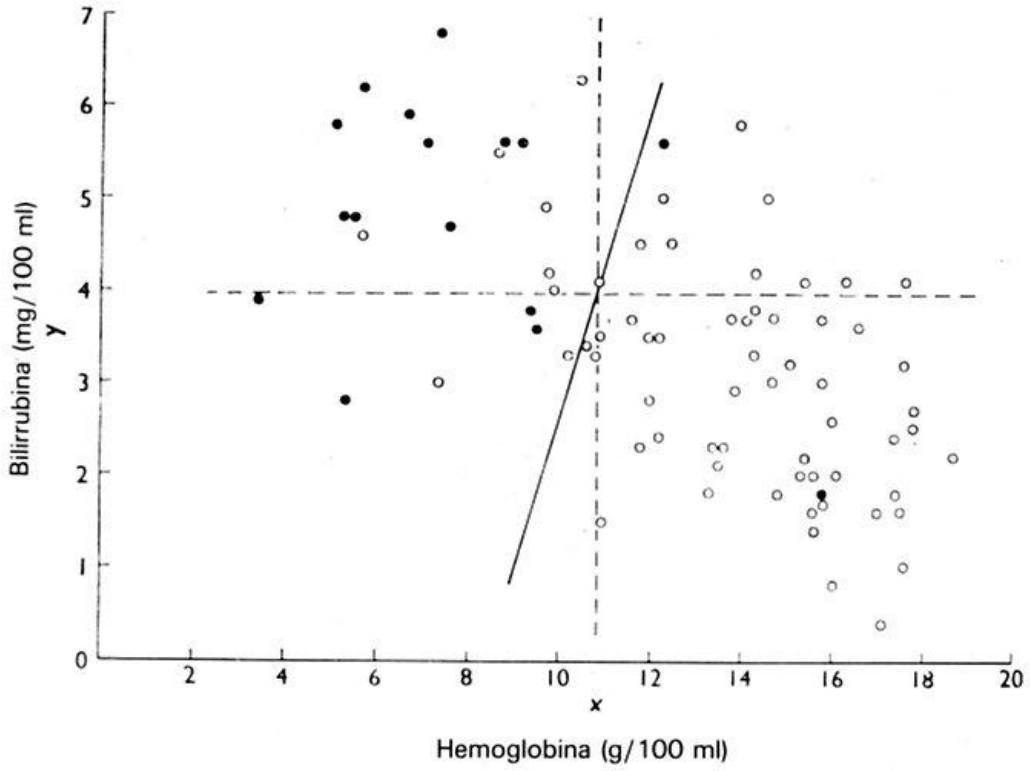
La función discriminante tiene como punto de corte 5,506, así si un neonato tiene un valor numérico dado por la función mayor a este valor, la probabilidad es más alta a sobrevivir y caso contrario a fallecer.

Así, si un neonato tiene un valor de hemoglobina de 16 y de Bilirrubina 3,2, aplicando la función discriminante obtenida:

Z = 0.6541 Hb - 0.3978 Bili

Z = 0.6541 (16) - 0.3978 (3,2)=9,227 como este valor numérico es superior a 5,506 el neonato tiene mayor probabilidad de vivir que de morir.

Esta situación puede ser más claramente apreciada en un diagrama de dispersión que muestre la función discriminante:



Observe que hay dos puntos negros del lado de los sobrevivientes

Éstos son dos neonatos mal clasificados por la función, pero en términos generales la función de discriminante se comporta bien.

Capítulo XI

Regresión Theil

(Regresión No Paramétrica)

La regresión Theil es una alternativa útil cuando no se cumplen las suposiciones en los cuales se sustenta el análisis de regresión lineal simple.

Theil propuso para el modelo clásico de regresión una forma muy particular de estimar los coeficientes alpha y beta:

Modelo de Regresión Theil

Sea el modelo de regresión clásico dado por:

$$\hat{Y} = \alpha + \beta X + \varepsilon$$

Donde las X son valores conocidos, α y β son parámetros desconocidos y Y es un valor observado de la variable aleatoria continua.

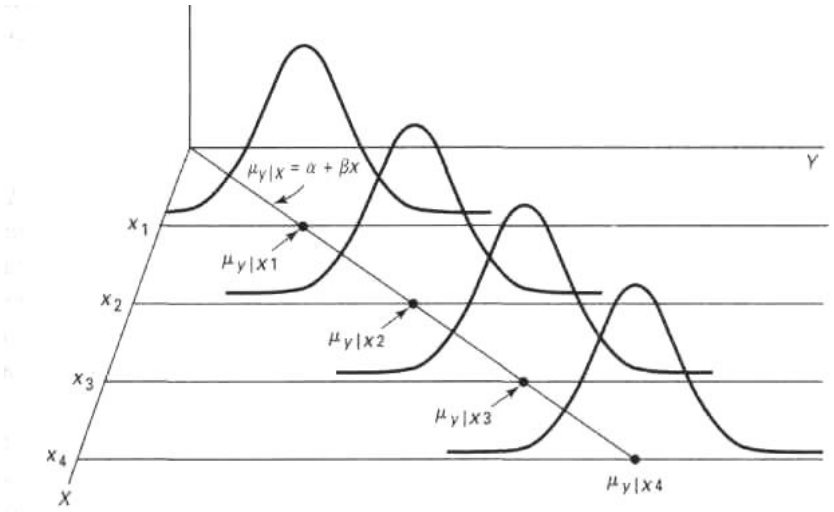
Para cada valor de X, se supone una subpoblación de Y valores, las ε

(epsilon) son errores aleatorios asociados a Y; y se consideran mutuamente independientes. Los epsilon se obtienen de:

$$\varepsilon = Y - (\alpha + \beta X)$$

Las X son todas distintas (no existen valores iguales) y no se repiten los mismos valores, y se tiene que X1<X2<X3<…<Xn.

Los datos se componen de n pares de observaciones de la muestra, (X1,Y1), (X2,Y2),…,(XnYn), donde el i-ésimo par representa las mediciones tomadas de la i-ésima unidad de asociación.



Estimación de los Parámetros Alpha y Beta

Estimador β:

Para obtener el estimador de Theil para beta, primero se forman todas las pendientes posibles de la muestra:

$$S_{ij} = \frac{(Y_j - Y_i)}{(X_j - X_i)}$$

Donde i<j. De esta forma existirán $N = C_n^2$ valores de S_{ij}.

El estimador de la pendiente β, es la mediana de los valores S_{ij}.

Esto es,

$\beta = mediana(\ S_{ij}\ )$

Estimador α:

El valor del intercepto se estima como una mediana de las diferencias de $\overline{Y} - \beta\overline{X}$, es decir:

$\alpha = mediana\ (\overline{Y} - \beta\overline{X})$

El estimador $\alpha = mediana\ (\overline{Y} - \beta\overline{X})$ se recomienda cuando el investigador no se inclina a suponer que los términos de error se distribuyen en forma simétrica alrededor de cero, que es lo más frecuente.

Ejemplo de Aplicación

En la tabla que sigue se dan los niveles (Y) plasmáticos de testosterona (nanogramos por mililitro) y los niveles de ácido cítrico seminal (mg/ml) en una muestra de 8 hombres adultos.

Se pretende estudiar la relación entre X e Y a través de la Regresión Theil.

Testosterna	230	175	315	290	275	150	360	425
Acido cítrico	421	278	618	482	465	105	550	750

Estimemos beta:

Las combinaciones de pares posibles de S_{ij} son:

$N = C_n^2 = 28$ valores ordenados de S_{ij}

-0,6618	**0,3846**	0,5637	1,0294
0,1445	0,4118	0,5927	
0,1838	0,4264	0,6801	
0,2532	0,4315	0,8333	
0,2614	0,4719	0,8824	
0,3216	0,5037	0,9836	
0,325	0,5263	1,0000	
0,3472	0,5297	1,0078	
0,3714	0,5348	1,0227	

El $S_{12} = \dfrac{175-230}{278-421} = +0{,}3846$

La mediana de los S_{ij} es 0,4878, es decir:

$$\beta = \frac{0{,}4719 + 0{,}5037}{2} = 0{,}4878$$

En consecuencia, la estimación del coeficiente de la pendiente beta es 0,4878.

Ahora obtengamos el estimador Alpha:

Calculando cada uno de los términos $\overline{Y} - \beta\overline{X}$. Desde luego se forman 36 pares ordenados:

13,5396	34,21	43,7823	53,6615	61,7086	75,43
19,0879	36,3448	47,136	54,8804	65,5508	76,8307
24,6362	36,4046	48,173	56,1603	69,0863	78,9655
26,4656	39,3916	49,2708	57,0152	69,9415	91,71
30,8563	39,7583	51,5267	58,1731	73,2952	95,2455
32,0139	41,8931	52,6248	59,15	73,477	98,781

La mediana de la serie ordenada anterior es:

$\alpha = mediana\ (\overline{Y} - \beta\overline{X})$=51,5267

Luego el modelo de regresión Theil ajustado es:

$$Testosterona = 51{,}5267 + 0{,}4878 Cítrico$$

Entonces se puede afirmar que los niveles de testosterona se incrementan en 0,4878 cuando el nivel de ácido cítrico varía de uno en uno.

Capítulo XII

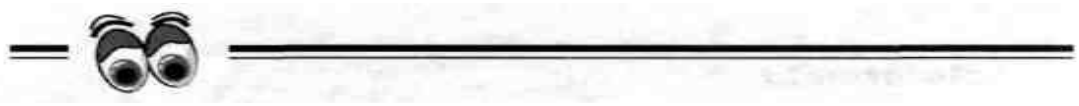
Regresión con Variables Ficticias

Conceptos y Definiciones

La regresión con variables ficticias (variables dummy) surge por la necesidad que tiene el investigador de involucrar variables cualitativas (o de atributos, o de categorías) en un análisis de regresión sea este simple o múltiple.

En algunas ocasiones el investigador maneja variables como:

- Estado civil (soltero, casado, viudo, divorciado)
- Sexo o género (masculino, femenino)
- Diagnóstico
- Grupo racial (blanco, negro, amarillo)
- Ocupación (sin y con trabajo)
- Zona de residencia (urbano, rural, suburbano)
- Tabaquismo (fumador cotidiano, exfumador, no fumador)
- Peso (muy pesado, medio pesado, poco pesado)
- Religión (católico, testigo, musulman, evangélico)
- Estatura (bajo, mediano, alto)
- Presión sanguínea (hipotenso, normotenso, hipertenso)
- Desempeño (bajo, medio, alto)
- Clima organizacional (favorable, desfavorable, aceptable)



En estos casos, el investigador se esfuerza por la inclusión de una o más de ellas porque sospecha un grado de contribución importante al reducir la suma de cuadrados del error y, por lo tanto, a

proporcionar estimaciones más precisas (de menor error estándar) de los parámetros de interés.

Las variables imaginarias (variables falsas o dummy) para poder incorporarlas en un modelo de regresión deben ser codificadas convenientemente. *La regla es introducir tantas variables imaginarias como categorías menos uno tenga la variable cualitativa, es decir, si una variable cualitativa tiene K categorías se introducirán en el modelo de regresión K-1 variables falsas*.

Una variable falsa es una variable que sólo toma un número finito de valores (como cero y uno) para identificar las diferentes categorías de una variable cualitativa.

Esta regla sólo es aplicable a aquellos casos en los cuales la ecuación de regresión tiene una constante o intercepto.

Técnicas de codificación

Para el ejemplo de tabaquismo que se refiere al principio si las categorías son: fumador, ex fumador (no ha fumado por 5 años o menos), ex fumador (no ha fumado por más de 5 años), no fuma. Como existen 4 categorías tendrán que crearse 3 variables falsas así:

X1= (1: para fumador, 0 para otro caso)

X2= (1: para ex fumador <=5 años; 0: para otro caso)

X3= (1: para ex fumador > 5 años; 0: para otro caso)

En otros casos (Gujarati,1997) , la codificación se puede establecer así:

Sea por ejemplo, el estudio del precio por onza de cola en función del tipo de almacén (descuento, cadena o de conveniencia), producto de marca o sin marca, llenado del envase.

Para almacén crea una sola variable D1 (dummy 1) que se codificará con 001 (si es un almacén de descuento), con 010 (si es almacén de cadena) y con 001 (si es almacén de conveniencia).

Para producto de marca crea una sola variable D2 (dummy 2) que codificará con 10 (si es un producto de marca) y como 01 (si es un producto sin marca).

Para el llenado del envase codificará así:

D3= 0001 (botella de 2 litros o 67,6 onzas)

= 0010 (botella de un litro o 28-33,8 onzas)

= 0100 (botella de 16 onzas)

= 1000 (latas de 12 onzas)

El comentario acerca de esta codificación plantea dos reflexiones una es que ocasiona la misma magnitud de disminución del error estándar del estimador, lo cual es favorable, pero por otro lado los resultados son más difíciles de interpretar. Con esta codificación también se tienen que crear menos variables y esto es una economía en el análisis. Si codificamos por la primera forma tendríamos que haber creado para el llenado del envase tres variables dummy porque tiene 4 categorías.

El comentario final sería codificar con 1 (uno) la categoría de interés y como 0 (cero) la otras. Esto facilitará la interpretación de los resultados.

Modelos de Regresión con variables falsas

Con tres variables (ejemplo del precio botellas de cola)

$$P = b_0 + b_1 D1 + b_2 D2 + b_3 D3$$

Donde:
P: Precio
D1= tipo de almacén
D2= marca del producto
D3= Llenado del envase

Con dos variables pero hay interacción entre ellas:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3(X1 * X2)$$

Término de interacción

Con términos polinómicos

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_1^2 + b_4X_2^2$$

Usos y aplicaciones de la Regresión con variables ficticias

(a) Para evaluar el efecto de variables cualitativas independientes

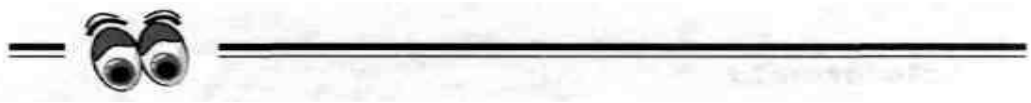
(b) Para desestacionalizar series de tiempo

(c) Para evaluar efectos de interacción entre variables independientes

(d) En casos de control estadístico del error (ANCOVA) que son modelos de regresión que contienen una mezcla de variables cuantitativas y cualitativas. Pero previamente deben probarse los supuestos de independencia entre la covariable y el tratamiento, también la homogeneidad de las pendientes y por último, la relación lineal entre la covariable y la variable respuesta.

Ejemplo de aplicación

Un grupo de investigadores en salud mental desea comparar tres métodos (A,B,C) para el tratamiento de la depresión grave. También se desea estudiar la relación entre la edad y la eficacia del tratamiento, así como la interacción (si existe) entre edad y tratamiento. Cada individuo de una muestra aleatoria simple de 36 pacientes, todos los cuales presentaban un diagnóstico y grado de depresión semejantes, recibió uno de los tres tratamientos. Los resultados se muestran a continuación.



La variable dependiente representa la eficacia del tratamiento (Y), la variable cuantitativa independiente X1 representa la edad del paciente, y la variable independiente cualitativa se refiere al tipo de tratamiento recibido que tiene tres niveles.

Se utiliza el siguiente código de variables ficticias para cuantificar la variable cualitativa:

Como ya tenemos una variable X1 que es la edad, se procede a crear una variable X2 y X3 para codificar el tratamiento, así:

X2= (1 si es el tratamiento A, 0 en otro caso)

X3= (1 si es el tratamiento B, 0 en otro caso)

Los términos de interacción se generan con la opción DATA seguido de TRANSFORMATIONS escribiendo:

IF TRAT=’A’ THEN X2=1 ELSE X2=0 (así creamos X2, recibir el tratamiento A)

IF TRAT=’B’ THEN X3=1 ELSE X3=0 (así creamos X3, recibir el tratamiento B)

Usando las mismas opciones del Menú, se crea la variable X4=X1*X2 (que representa la interacción recibir A con la edad) y la interacción X5 =X1*X3 (que representa la interacción de la edad y recibir el tratamiento B). El tratamiento C queda representado por el intercepto.

Matriz de Datos

Una vez creadas las variables se procede a alimentar a la computadora con nuestros datos:

Medida de eficacia Y	Edad X1	Método de Tratamiento
56	21	A
55	28	B
63	33	B
52	33	C
58	38	A
65	43	C
64	48	B
61	53	C
69	53	B
73	58	A
62	63	A
70	67	C
41	23	C
40	30	B
46	33	A
48	42	C
45	43	B
58	43	C
55	45	A
57	48	B
62	58	B
47	29	C
64	66	A
60	67	A
28	19	B
25	23	C
71	67	A
62	56	B
50	45	A
46	37	B
34	27	B
59	47	A
36	29	C
71	59	C
62	51	A
71	63	C

MATRIZ DE DATOS CODIFICADOS QUE GENERÓ EL SOFTWARE TRAS LA ALIMENTACIÓN DE LAS TRES VARIABLES

Y	X1	X2	X3	X1X2	X1X3
56	21	1	0	21	0
55	28	1	0	28	0
63	33	1	0	33	0
52	33	1	0	33	0
58	38	1	0	38	0
65	43	1	0	43	0
64	48	1	0	48	0
61	53	1	0	53	0
69	53	1	0	53	0
73	58	1	0	58	0
62	63	1	0	63	0
70	67	1	0	67	0
41	23	1	1	0	23
40	30	0	1	0	30
46	33	0	1	0	33
48	42	0	1	0	42
45	43	0	1	0	43
58	43	0	1	0	43
55	45	0	1	0	45
57	48	0	1	0	48
62	58	0	1	0	58
47	29	0	1	0	29
64	66	0	1	0	66
60	67	0	1	0	67
28	19	0	1	0	0
25	23	0	1	0	0
71	67	0	1	0	0
62	56	0	1	0	0
50	45	0	1	0	0
46	37	0	1	0	0
34	27	0	1	0	0
59	47	0	0	0	0
36	29	0	0	0	0
71	59	0	0	0	0
62	51	0	0	0	0
71	63	0	0	0	0

Al examinar la salida impresa de los resultados se obtiene mayor información acerca de la naturaleza de las relaciones entre las variables:

```
Statistix 8.0

Unweighted Least Squares Linear Regression of Y

Predictor
```

Variables	Coefficient	Std Error	T	P	VIF
Constant	24.0453	4.50528	5.34	0.0000	
X1	0.75214	0.08286	9.08	0.0000	1.8
X2	15.3884	4.83035	3.19	0.0034	6.9
X3	-7.62024	3.02533	-2.52	0.0173	2.9
X4	-0.25718	0.10507	-2.45	0.0204	7.2
X5	0.01722	0.05622	0.31	0.7615	2.0

R-Squared	0.8430	Resid. Mean Square (MSE)	28.2315
Adjusted R-Squared	0.8168	Standard Deviation	5.31333

Source	DF	SS	MS	F	P
Regression	5	4548.06	909.611	32.22	0.0000
Residual	30	846.94	28.231		
Total	35	5395.00			

Cases Included 36 Missing Cases 0

Como se aprecia, la ecuación por mínimos cuadrados es:

$$\hat{Y} = 24{,}04 + 0{,}75X1 + 15{,}4X2 - 7{,}62X3 - 0{,}26X4 + 0{,}02X5$$

Cuyo R-cuadrado es:

$$R^2 = 0{,}84$$

Lo cual indica que el 84% de la variación en la eficacia de los tratamientos se explican por la edad, el tratamiento y sus interacciones.

Las tres ecuaciones de regresión para los tres tratamientos son las siguientes:

Para el Tratamiento A

$$\hat{Y} = (24{,}04 + 15{,}38) + (0{,}75 - 0{,}257)X1$$

Para el Tratamiento B

$$\hat{Y} = (24{,}04 - 7{,}62) + (0{,}75 + 0{,}017)X1$$

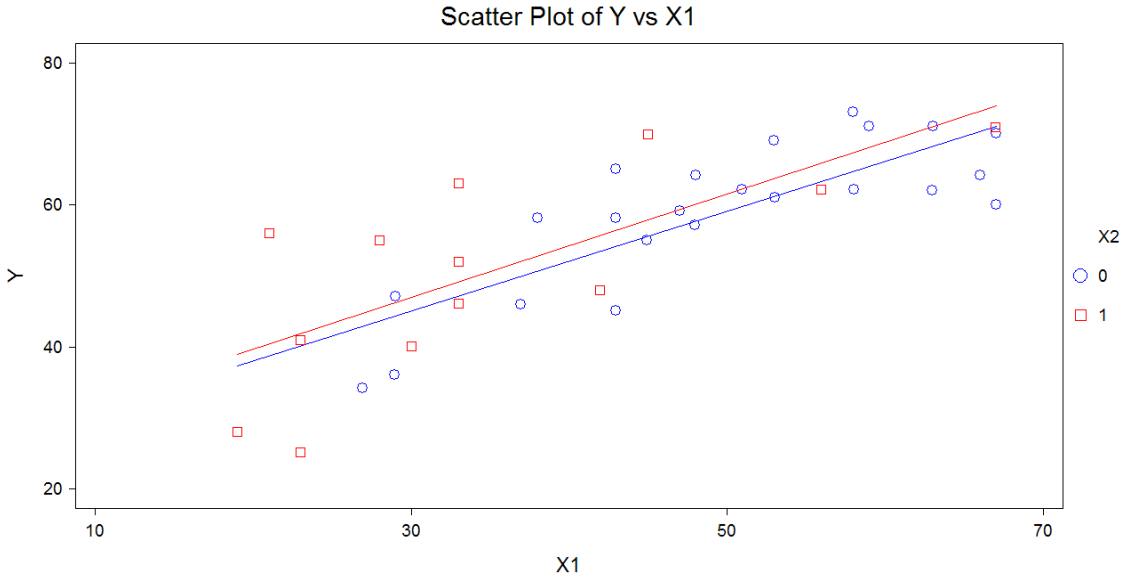
Para el Tratamiento C

$$\hat{Y} = 24{,}04 + 0{,}75X1$$

La conclusión es:

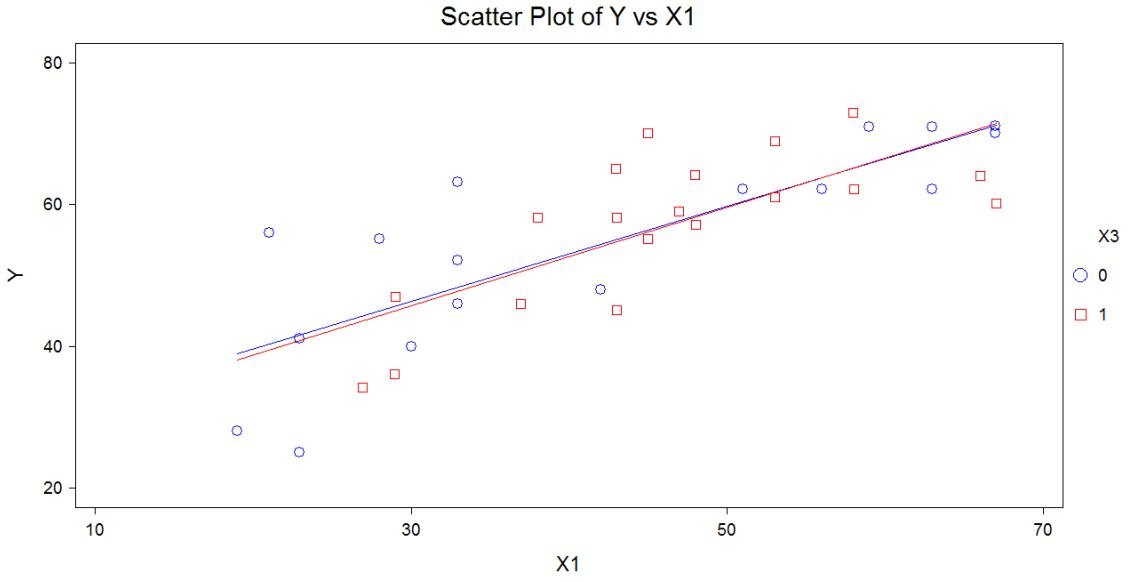
Todos Los efectos fueron significativos

La gráfica del diagrama de dispersión de la eficacia versus la edad muestra que las pendientes no son iguales:



Así mismo, la gráfica de la eficacia versus edad para el tratamiento B, revelan que las pendientes no son iguales:

E



Estas gráficas justifican o avalan la existencia de una interación significativa (al entrecruzarse) entre el tratamiento y la edad.

Ejemplo 1

Estadística descriptiva

Número de horas gastadas en ver televisión por 180 estudiantes durante una semana

ALUMNO	HORAS	ALUMNO	HORAS	ALUMNO	HORAS	ALUMNO	HORAS	ALUMNO	HORAS
1	23	37	23	73	17	109	21	145	18
2	17	38	20	74	23	110	17	146	19
3	23	39	23	75	22	111	19	147	17
4	18	40	20	76	18	112	8	148	21
5	21	41	9	77	20	113	20	149	16
6	22	42	21	78	8	114	16	150	18
7	18	43	17	79	22	115	19	151	17
8	20	44	22	80	20	116	17	152	21
9	24	45	22	81	21	117	25	153	20
10	22	46	20	82	18	118	19	154	15
11	19	47	19	83	24	119	18	155	19
12	18	48	23	84	19	120	22	156	22
13	21	49	20	85	21	121	19	157	25
14	20	50	19	86	22	122	24	158	20
15	20	51	25	87	19	123	20	159	19
16	17	52	18	88	23	124	17	160	21
17	18	53	22	89	19	125	21	161	18
18	20	54	19	90	21	126	15	162	20
19	18	55	21	91	17	127	23	163	22
20	21	56	20	92	24	128	17	164	19
21	17	57	24	93	21	129	17	165	20
22	20	58	20	94	17	130	20	166	16
23	7	59	19	95	23	131	24	167	19
24	20	60	20	96	19	132	16	168	21

25	23	61	23	97	22	133	17	169	24
26	18	62	18	98	17	134	22	170	20
27	18	63	25	99	20	135	16	171	19
28	21	64	21	100	16	136	21	172	21
29	19	65	20	101	17	137	19	173	23
30	19	66	17	102	17	138	17	174	19
31	15	67	25	103	15	139	16	175	21
32	21	68	20	104	21	140	20	176	20
33	25	69	25	105	23	141	16	177	23
34	22	70	2!	106	16	142	18	178	25
35	24	71	18	107	18	143	16	179	18
36	21	72	22	108	18	144	22	180	24

Ejemplo 2

Pesos en libras de 150 adultos

165	179	168	178	175	176	181	160	176	177	176	176	189	178
161	181	163	182	177	166	184	179	178	176	188	165	179	190
172	165	193	159	179	174	173	171	179	178	169	183	175	177
165	189	175	176	186	184	186	162	176	168	184	191	179	160
163	167	176	186	175	171	171	187	176	186	177	183	180	181
171	176	175	187	170	167	184	173	175	169	173	170	174	177
163	179	187	168	178	178	170	175	173	181	184	180	162	178
188	165	177	172	168	186	174	180	189	168	188	181	177	175
166	179	188	185	178	176	177	179	176	180	183	184	176	182
188	186	170	178										

Ejemplo 3

Altura en pies de 100 árboles de una finca

2	2	9	5	9	13	10	16	19	20	3	2	8	5	8
13	11	17	17	24	1	2	9	9	5	14	11	17	15	23
4	2	8	5	5	14	13	16	16	24	4	4	8	7	6
10	10	15	16	21	1	4	9	6	9	11	13	15	22	28
3	1	9	6	*5*	13	13	16	20	26	1	1	8	8	8

12 13 19 21 28 3 6 6 5 13 14 18 15 21 29
3 5 7 8 13 12 16 16 24 26

Ejemplo 4

En un experimento en psicología, se pide a varios sujetos memorizar cierta secuencia de palabras. Los datos representan el tiempo (en segundos) que cada sujeto tardó en memorizarlas:

100 107 34 57 66 30 79 84 118 77 135 95 130 138 52 126
89 128 100 88 61 108 79 37 93 116 45 57 112 73 129 46
107 109 32 106 122 41 70 96 98 117 97 99 62 88 85 149
75 105 50 99 50 79 43 90 114 53 123 100 69 87 64 85
126 100 102 112 78 118 135 110 64 62 107 127 102 129 88 123
98 110 93 135 58 73 80 125 88 142 103 149 90 145 96 146
119 76 93 99

Ejemplo 5

Medición de la presión sistólica de 100 adultos que se presentaron para un examen físico antes de un empleo:

104 112 128 139 118 132 132 112 106 107 129 125 103 125 104 129
126 126 115 118 117 116 113 122 123 107 122 105 110 133 142 143
116 114 129 117 106 124 115 118 123 101 123 121 124 120 116 117
105 120 146 121 120 102 138 106 113 130 111 123 124 120 113 115
114 122 116 108 122 112 112 123 116 115 116 111 120 119 122 123
124 111 121 111 114 123 107 120 120 106 118 116 135 121 123 117
124 122 134 131 158 176 179 177 158 180 175 168 171 176 178 174
189 171 164 185 180 172 171 181 170

Ejemplo 6

Dos técnicos de laboratorio, A y B, determinaron la cantidad de hemoglobina en 15 muestras de sangre. Los resultados se expresan en gramos por 100 cc de sangre:

	Técnico			Técnico	
No. de la muestra	*A*	*B*	No. de la muestra	*A*	*B*
1	15.38	15.7 1	9	12.47	12.74
2	17.78	17.40	10	12.95	13.78
3	16.77	16.94	II	11.28	11.65

4	16.05	16.75	12	10.65	10.08
5	17.67	16.24	13	10.80	10.15
6	13.16	13.85	14	15.70	14.92
7	13.42	12.02	15	12.23	12.27
8	18.85	18.64			

Ejemplo 7

Se registraron las pulsaciones por minuto de 12 sujetos antes y después de haber fumado una cantidad uniforme de marihuana:

Pulsaciones

Sujeto	1	2	3	4	5	6	7	8	9	10	11	12
Antes	75	61	62	68	58	70	59	79	68	80	64	75
Después	82	70	74	80	65	80	70	88	77	90	75	87

Ejemplo 8

Niveles de cierto producto químico en la sangre de 15 sujetos antes y después que han sido sometidos a una situación que produce ansiedad:

Nivel del producto químico

Sujeto	1	2	3	4	*5*	6	7	8	9	10	11	12	13	14	15
Antes(B)	9	17	14	8	15	20	18	12	10	22	18	7	14	7	20
Después(A)	10	22	24	28	10	15	14	12	21	25	22	11	16	10	27
Diferencia(A-B)	+1	+5	+10	+20	-5	-5	- 4	0	+11	+3	+4	+4	+2	+3	+7

Ejemplo 9

Un psicólogo industrial deseaba hacer un estudio sobre los efectos que producían 5 diferentes incentivos (1. Dinero, 2.Cambio de turno o Departamento, 3. Tiempo Libre Remunerado, 4. Tiempo adicional para el café, 5. Recompensa especial) . Se llevó a cabo un experimento en un departamento de cada uno de los 5 talleres de una gran empresa. . Tomaron parte en el experimento los empleados del turno de las 11:00 P.M- a las 7:00 A.M. De cada taller participaron 10 empleados. Se les prometieron 5 tipos de recompensas si aumentaban la producción durante la semana del experimento.

La variable respuesta fue la diferencia entre el número de artículos producidos durante la semana del experimento y el número promedio de artículos producidos por semana durante el mes anterior al experimento, para cada uno de los empleados y cada uno de los tratamientos.

RECOMPENSA (Tratamiento)

TRAT1	TRAT2	TRAT3	TRAT4	TRAT5
76	52	37	19	11
70	52	26	21	15
59	43	28	16	23
77	48	38	23	15
59	43	25	23	25
69	56	30	23	18
80	52	28	14	20
78	53	25	15	18
61	58	26	13	20
66	50	25	16	16

Total	695	507	288	183	181	1854
Media	69.5	50.7	28.8	18.3	18.1	37.08
Varianza	65.17	24.23	23.73	15.79	16.99	438.65

Ejemplo 10

Diferencia entre los niveles de depresión obtenidos por 12 pacientes antes y después de que recibieran cuatro drogas diferentes.

Drogas

Duración de la enfermedad	A	B	C	D	Total	Media
Corta	12	11	16	15	54	13.50
Mediana	10	13	15	17	55	13.75
Larga	8	8	10	10	36	9.00
Total	30	32	41	42	145	
Media	10.00	10.67	13.67	14.00		12.08

Los investigadores, que deseaban eliminar los efectos de la duración de la enfermedad, seleccionaron al azar cuatro pacientes de cada una de tres categorías de duración de la enfermedad

— corta, mediana, larga — y luego asignaron al azar cada uno de los pacientes de los grupos de duración de la enfermedad (bloque) a los cuatro tratamientos o drogas. La medida que tomaron fue la diferencia entre los puntajes obtenidos por los pacientes antes y después de la aplicación de la droga en una prueba para medir el nivel de depresión.

Ejemplo 11

A continuación se dan codificados los datos referentes al coeficiente intelectual de 15 niños de un barrio pobre y los puntajes de ajuste personal.

Puntaje de ajuste, Y	4	5	4	6	5	7	8	9	13	11	15	14	13	16	17
CI, X	-10	-8	-7	-5	-4	0	3	5	8	10	12	14	15	16	20

Ejemplo 12

El número de horas por semana que gastaron diez universitarios estudiando y su promedio de puntaje de notas acumulativas es:

Promedio de notas ,Y	2.1	2.7	2.6	2.5	3.5	3.0	3.5	3.7	2.9	4.0
Horas de estudio, X	5	6	7	8	9	10	11	12	13	14

Ejemplo 13

A una muestra aleatoria de 12 estudiantes de quinto grado, se le administraron dos pruebas. Una tenía por objeto medir el nivel de hostilidad y la otra la comprensión de lectura. La Tabla muestra los puntajes (codificados):

Comprensión de lectura, Y	98	90	95	80	84	79	67	70	65	57	55	50
Puntaje de hostilidad, X	0	1	2	3	4	5	6	7	8	9	10	11

Ejemplo 14

Se recogieron en 10 comunidades de un país en vías de desarrollo.

Incidencia de bocio En la población, Y(%)	60	75	50	55	45	33	50	52	35	30
Contenido de yodo En el agua, X (microgramo/litro)	2	2	3	5	7	8	8	8	10	12

Ejemplo 15

La Tabla presenta una muestra aleatoria de 11 sujetos, el tiempo en minutos requerido para completar una tarea y el número de minutos gastados en aprender esa tarea:

Tiempo para Hacer la tarea, Y	40	35	20	38	17	26	28	22	12	12	5
Tiempo gastado en aprender, X	30	30	40	40	50	50	60	60	60	70	70

Ejemplo 16

Se seleccionó una muestra de 100 personas de la nómina de 5 organizaciones profesionales y se les preguntó cuándo habían decidido seguir la profesión:

La tabla que sigue muestra los resultados:

	Profesión					
Época de elección De profesión	*A*		*C*	*D*	*E*	Total
Antes del colegio	30	20	30	10	10	100
Durante el colegio	40	55	46	20	30	191
Durante la universidad	30	25	24	70	60	209
Total	100	100	100	100	100	500

Ejemplo 17

Se seleccionó una muestra al azar de 100 alumnos de último año de un colegio de cada uno de los siguientes tres grupos de rendimiento atlético: alto, medio y bajo. Los muchachos se clasificaron de acuerdo con la inteligencia tal como aparece en la Tabla. ¿Indican estos datos una diferencial en la distribución de la inteligencia entre los tres grupos? Sea alfa = 0.05.

	Rendimiento atlético			
Inteligencia	Alto	Medio	Bajo	Total
Alta	28	30	34	92
Media	40	38	42	120
Baja	32	32	24	88
Total	100	100	100	300

Ejemplo 18

Unos psiquiatras que estaban haciendo un estudio sobre el alcoholismo, seleccionaron una muestra aleatoria de pacientes sometidos a tratamientos en una clínica de alcoholismo, clasificada en tres clases sociales. Luego los clasificaron de acuerdo con su nivel de comunicación con el jefe de terapeutas. La Tabla muestra los resultados. ¿Indican estos datos una falta de homogeneidad entre las clases sociales respecto del nivel de comunicación con el jefe de terapeutas? Sea alfa = 0.05.

	Clase social			
Nivel de comunicación	1	II	III	Total
Bueno	45	60	68	173
Regular	10	15	25	50
Deficiente	5	15	32	52
Total	60	90	125	275

Ejemplo 19

Supongamos que estamos interesados en conocer el ecosistema del Lago de Maracaibo concretamente se desea comparar dos áreas de aproximadamente un cuarto de milla cuadrada; una de descarga de productos químicos y biológicos (Área B) y otra semejante (Área A) en cuanto a tamaño, profundidad y cambios estacionales, pero no es de descarga:

AREA *A* (Control)			*AREA* B (Descarga)		
Sitiol	Sitio2	Sitio3	Sitiol	Sitio2	Sitio3
18	19	18	21	19	19
16	*20*	18	20	20	23
16	19	20	18	21	21

Nota: la variable respuesta es la concentración de mercurio.

Se pide:

1. Escribir el modelo aditivo lineal para este ensayo.
2. Proponer el sistema hipotético.
3. Desarrollar el análisis de varianza.
4. Expresar las conclusiones.

Ejemplo 20

Conjunto de Datos sobre Bienes Raíces:

x_1 = Precio de venta en miles de dólares
x_2 = Número de recámaras
x_3 = Extensión de la casa en pies cuadrados
x_4 = Piscina (1 = sí, 0 = no)
x_5 = Distancia desde el centro de la ciudad
x_6 = Municipio
x_7 = Garaje anexo (1 = sí, 0 = no)
x_8 = Número de baños

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
263.115	4	2 349	0	17	5	1	2
182.385	4	2 102	1	19	4	0	2
242.055	3	2 271	1	12	3	0	2
213.57	2	2 188	1	16	2	0	2.5
139.86	2	2 148	1	28	1	0	1.5
245.43	2	2 117	0	12	1	1	2
327.24	6	2 484	1	15	3	1	2
271.755	2	2 130	1	9	2	1	2.5
221.13	3	2 254	0	18	1	0	1.5
266.625	4	2 385	1	13	4	1	2
292.41	4	2 108	1	14	3	1	2
208.98	2	1 715	1	8	4	1	1.5
270.81	6	2 495	1	7	4	1	2
246.105	4	2 073	1	18	3	1	2
194.4	2	2 283	1	11	3	0	2
281.34	3	2 119	1	16	2	1	2
172.665	4	2 189	0	16	3	0	2
207.495	5	2 316	0	21	4	0	2.5
198.855	3	2 220	0	10	4	1	2
209.25	6	1 901	0	15	4	1	2
252.315	4	2 624	1	8	4	1	2
192.915	4	1 938	0	14	2	1	2.5
209.25	5	2 101	1	20	5	0	1.5
345.33	8	2 644	1	9	4	1	2
326.295	6	2 141	1	11	5	1	3
173.07	2	2 198	0	21	5	1	1.5
186.975	2	1 912	1	26	4	0	2
257.175	2	2 117	1	9	4	1	2
233.01	3	2 162	1	14	3	1	1.5
180.36	2	2 041	1	11	5	0	2
233.955	2	1 712	1	19	3	1	2
207.09	2	1 974	1	11	5	1	2
247.725	5	2 438	1	16	2	1	2
166.185	3	2 019	0	16	2	1	2
177.12	2	1 919	1	10	5	1	2

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
182.655	4	2 023	0	14	4	0	2.5
216	4	2 310	1	19	2	0	2
312.12	6	2 639	1	7	5	1	2.5
199.8	3	2 069	1	19	3	1	2
273.24	5	2 182	1	16	2	1	3
206.01	3	2 090	0	9	3	0	1.5
232.2	3	1 928	0	16	1	1	1.5
198.315	4	2 056	0	19	1	1	1.5
205.065	3	2 012	0	20	4	0	2
175.635	4	2 262	0	24	4	1	2
307.8	3	2 431	0	21	2	1	3
269.19	5	2 217	1	8	5	1	3
224.775	3	2 157	1	17	1	1	2.5
171.585	3	2 014	0	16	4	0	2
216.81	3	2 221	1	15	1	1	2
192.645	6	2 236	0	14	1	0	2
236.385	5	2 189	1	20	3	1	2
172.395	3	2 218	1	23	3	0	2
251.37	3	1 937	1	12	2	1	2
245.97	6	2 296	1	7	3	1	3
147.42	6	1 749	0	12	1	0	2
176.04	4	2 230	1	15	1	1	2
228.42	3	2 263	1	17	5	1	1.5
166.455	3	1 593	0	19	3	0	2.5
189.405	4	2 221	1	24	1	1	2
312.12	7	2 403	1	13	3	1	3
289.845	6	2 036	1	21	3	1	3
269.865	5	2 170	0	11	4	1	2.5
154.305	2	2 007	1	13	2	0	2
222.075	2	2 054	1	9	5	1	2
209.655	5	2 247	0	13	2	1	2
190.89	3	2 190	0	18	3	1	2
254.34	4	2 495	0	15	3	1	2
207.495	3	2 080	0	10	2	0	2
209.655	4	2 210	0	19	2	1	2
294.03	2	2 133	1	13	2	1	2.5
176.31	2	2 037	0	17	3	0	2
294.3	7	2 448	1	8	4	1	2
223.965	3	1 900	0	6	1	1	2
125.01	2	1 871	1	18	4	0	1.5
236.8035	4	2 583.9	0	17	5	1	2
164.1465	4	2 312.2	1	19	4	0	2
217.8495	3	2 498.1	1	12	3	0	2
192.213	2	2 406.8	1	16	2	0	2.5
125.874	2	2 362.8	1	28	1	0	1.5
220.887	2	2 328.7	0	12	1	1	2

Ejemplo 21

Conjunto de Datos sobre la Organización para la Cooperación y el Desarrollo Económico:

x_1 = Precio de venta en miles de dólares
x_2 = Número de recámaras
x_3 = Extensión de la casa en pies cuadrados
x_4 = Piscina (1 = sí, 0 = no)
x_5 = Distancia desde el centro de la ciudad
x_6 = Municipio
x_7 = Garaje anexo (1 = sí, 0 = no)
x_8 = Número de baños

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
263.115	4	2 349	0	17	5	1	2
182.385	4	2 102	1	19	4	0	2
242.055	3	2 271	1	12	3	0	2
213.57	2	2 188	1	16	2	0	2.5
139.86	2	2 148	1	28	1	0	1.5
245.43	2	2 117	0	12	1	1	2
327.24	6	2 484	1	15	3	1	2
271.755	2	2 130	1	9	2	1	2.5
221.13	3	2 254	0	18	1	0	1.5
266.625	4	2 385	1	13	4	1	2
292.41	4	2 108	1	14	3	1	2
208.98	2	1 715	1	8	4	1	1.5
270.81	6	2 495	1	7	4	1	2
246.105	4	2 073	1	18	3	1	2
194.4	2	2 283	1	11	3	0	2
281.34	3	2 119	1	16	2	1	2
172.665	4	2 189	0	16	3	0	2
207.495	5	2 316	0	21	4	0	2.5
198.855	3	2 220	0	10	4	1	2
209.25	6	1 901	0	15	4	1	2
252.315	4	2 624	1	8	4	1	2
192.915	4	1 938	0	14	2	1	2.5
209.25	5	2 101	1	20	5	0	1.5
345.33	8	2 644	1	9	4	1	2
326.295	6	2 141	1	11	5	1	3
173.07	2	2 198	0	21	5	1	1.5
186.975	2	1 912	1	26	4	0	2
257.175	2	2 117	1	9	4	1	2
233.01	3	2 162	1	14	3	1	1.5
180.36	2	2 041	1	11	5	0	2
233.955	2	1 712	1	19	3	1	2
207.09	2	1 974	1	11	5	1	2
247.725	5	2 438	1	16	2	1	2
166.185	3	2 019	0	16	2	1	2
177.12	2	1 919	1	10	5	1	2
182.655	4	2 023	0	14	4	0	2.5
216	4	2 310	1	19	2	0	2
312.12	6	2 639	1	7	5	1	2.5
199.8	3	2 069	1	19	3	1	2
273.24	5	2 182	1	16	2	1	3
206.01	3	2 090	0	9	3	0	1.5
232.2	3	1 928	0	16	1	1	1.5
198.315	4	2 056	0	19	1	1	1.5
205.065	3	2 012	0	20	4	0	2
175.635	4	2 262	0	24	4	1	2
307.8	3	2 431	0	21	2	1	3
269.19	5	2 217	1	8	5	1	3
224.775	3	2 157	1	17	1	1	2.5
171.585	3	2 014	0	16	4	0	2
216.81	3	2 221	1	15	1	1	2
192.645	6	2 236	0	14	1	0	2
236.385	5	2 189	1	20	3	1	2
172.395	3	2 218	1	23	3	0	2
251.37	3	1 937	1	12	2	1	2
245.97	6	2 296	1	7	3	1	3
147.42	6	1 749	0	12	1	0	2
176.04	4	2 230	1	15	1	1	2
228.42	3	2 263	1	17	5	1	1.5
166.455	3	1 593	0	19	3	0	2.5
189.405	4	2 221	1	24	1	1	2
312.12	7	2 403	1	13	3	1	3
289.845	6	2 036	1	21	3	1	3
269.865	5	2 170	0	11	4	1	2.5
154.305	2	2 007	1	13	2	0	2
222.075	2	2 054	1	9	5	1	2
209.655	5	2 247	0	13	2	1	2
190.89	3	2 190	0	18	3	1	2
254.34	4	2 495	0	15	3	1	2
207.495	3	2 080	0	10	2	0	2
209.655	4	2 210	0	19	2	1	2
294.03	2	2 133	1	13	2	1	2.5
176.31	2	2 037	0	17	3	0	2
294.3	7	2 448	1	8	4	1	2
223.965	3	1 900	0	6	1	1	2
125.01	2	1 871	1	18	4	0	1.5
236.8035	4	2 583.9	0	17	5	1	2
164.1465	4	2 312.2	1	19	4	0	2
217.8495	3	2 498.1	1	12	3	0	2
192.213	2	2 406.8	1	16	2	0	2.5
125.874	2	2 362.8	1	28	1	0	1.5
220.887	2	2 328.7	0	12	1	1	2

Ejemplo 22

Experimento Factorial A x B x C:

Se realizó una investigación en tres planteles (Privado: Al, Subvencionado por el estado, pero administrado por particulares: A2, y Público: A3) para determinar la influencia en el aprendizaje de la matemática de 4 métodos basados en teorías (Piagetiana. 131, Conductista: B2, Gestaltista B3 y teoría personal del investigador de carácter lúdico: B4). Tanto los planteles como los niños fueron se1eccionados al azar (5 niños por maestro y por método). Cinco tipos de Maestros (C1 sin capacitación, C2= maestro, C3= técnico, C4= licenciado y C5 profesor de matemática) fueron preparados con anterioridad sobre las bases teóricas de los métodos y procedimientos a seguir en la aplicación de los mismos. Los métodos se aplicaron a niños de 7 años que cursaban el primer nivel de educación básica. La variable respuesta es la nota obtenida en un examen en la escala tradicional del cero al veinte:

Al				A2				A3			
B1	B2	B3	134	131	132	133	B4	131	132	133	134
4	3 4	0	4	2	8	2	5	8	7	3	
2	55	0	5	1	6	1	7	6	2	1	
3	23	1	7	3	9	0	6	7	3	2	Cl
1	12	1	8	2	6	1	4	5	4	1	
7	01	2	6	0	7	1	5	4	3	2	
3	13	0	3	4	5	2	8	6	2	1	
2	02	2	2	2	3	0	3	3	3	1	
5	25	3	4	3	4	3	5	2	7	0	C2
4	37	2	5	2	5	1	4	4	5	1	
0	04	1	1	1	3	0	6	5	3	2	
1	15	0	6	2	5	1	8	2	2	3	
3	38	1	3	3	4	3	4	5	3	1	
4	26	1	2	2	3	2	3	3	4	1	C3
2	19	2	4	0	3	0	6	7	5	2	
8	05	0	1	1	2	2	5	4	6	2	
3	27	1	3	2	4	3	8	5	2	1	
10	3 9	3	2	1	3	1	7	3	3	1	
9	68	2	1	2	5	0	7	3	3	1	C4
11	8 7	2	4	1	4	2	4	6	5	2	
12	79	3	1	3	3	1	5	4	4	0	
20	7 8	2	5	2	5	1	‘6	5	8	0	
18	9 4	1	8	4	3	0	3	8	3	1	
15	11 11	6	6	3	4	2	5	2	5	0	C5
17	13 6	3	4	2	7	0	3	2	5	0	
14	16 10	1	5	2	9	1	4	0	6	3	

Se pide:

(a) Plantear las hipótesis.
(b) Realizar el análisis de varianza.
(c) Cuál es la mejor combinación de factores (A, B, C) que logra el mejor rendimiento?

(d) Representar gráficamente (gráfico de tendencias) y verificar la existencia de interacción.
(e) Expresar las conclusiones que se puedan derivar del análisis.

Ejemplo 23

A continuación se muestra una tabla con los datos de los 36 casos que realizaron el experimento.

Caso	Tipo de música	Luz	Droga	Rendimiento en cálculo matemático
1	Heavy	Natural	Tratam.	11
2				13
3				15
4			Placebo	9
5				12
6				14
7		Artificial	Tratam.	5
8				7
9				8
10			Placebo	5
11				7
12				8
13	Ambiental	Natural	Tratam.	15
14				16
15				18
16			Placebo	11
17				14
18				13
19		Artificial	Tratam.	9
20				10
21				13
22			Placebo	4
23				6
24				7
25	Mozart	Natural	Tratam.	19
26				23
27				19
28			Placebo	14
29				13
30				12
31		Artificial	Tratam.	13
32				17
33				14
34			Placebo	9
35				7
36				7

Los valores de las variables independientes son:

- Tipo (1:heavy, 2: ambiental, 3: mozart).
- Luz (1: natural, 2: artifical).
- Droga (1: tratamiento, 2: placebo).

Como existen tres variables independientes se debería realizar un anova con las tres vi, pero como demostración del comando UNIANOVA primero se va a realizar un anova con dos vi (tipo de música y luz) y posteriormente se añadirá la tercera vi (droga).

Ejemplo 24

En una investigación realizada en la ciudad de Caracas se desea determinar si existe relación entre la variable ausencia del padre en el hogar y el nivel de autoestima alcanzado por los hijos varones. Se supone que la variable clima familiar influye también sobre la autoestima y es probable que un

mejor o peor clima familiar se refleje también en una autoestima alta o baja. La variable independiente (ausencia del padre) tiene tres niveles de familias: con padrasto, con padre y sin padre. La variable dependiente (autoestima) y la Covariable (clima) se miden en puntajes. La matriz de datos obtenida es:

Grupos (Tratamientos)

A(Flias. Con Padrasto)		B(Flias.Con Padre)		C(Flias. Sin Padre)	
Autoestima	Clima	Autoestima	Clima	Autoestima	Clima
Y	X	Y	X	Y	X
15	30	25	28	5	10
10	20	10	12	10	15
5	15	15	20	20	20
10	20	15	10	5	10
20	25	10	10	10	10

Ejemplo 25

Se desea saber si existe relación entre el estilo de liderazgo que utiliza el docente en el aula y el rendimiento académico de los alumnos. Sin embargo, se sabe que el rendimiento puede estar afectado por el tipo de materia o contenido que trabajan y por los recursos pedagógicos que utiliza el docente, por lo que se desea controlar estas dos variables extrañas. La matriz de datos es:

Recursos	Materia o Contenidos		
	Matemática	Biología	Castellano
Pizarra	8, 10, 9 A	16, 16, 15 D	11, 9, 10 P
Audiovisuales	18, 16, 17 D	13, 11, 12 P	10, 9, 11 A
Modelos	13, 12, 14 P	13, 11, 12 A	20, 18, 19 D

Siendo (A = autoritario, D = democrático y C = lesefer o permisivo) los estilos de liderazgos del docente en el aula.

Ejemplo 26

Un investigador desea saber si existe relación entre el tipo de autoridad familiar que ha ejercido en el hogar el padre sobre sus hijos y el nivel de autoestima de los hijos; para ello se ha aplicado un instrumento obteniéndose los resultados siguientes:

Tipo de autoridad	Nivel de auto estima de los hijos		
Familiar	Alta	Media	Baja
Autocrática	2	8	15

Democrática 20 11 4

Ejemplo 27

Una investigación sociológica realizada en Caracas deseaba determinar la relación entre la ausencia del padre en el hogar (TRAT) y el nivel de autoestima (Y) alcanzado por los hijos varones. Pero sospecha que esta relación pudiera estar afectada por el clima familiar (X=covariable): La matriz de datos original es:

Familias con Padrastro		Familias con Padre		Familias sin Padre	
Y	X	Y	X	Y	X
15	30	25	28	05	10
10	20	10	12	10	15
05	15	15	20	20	20
10	20	15	10	05	10
20	25	10	10	10	10

Se pide:

(1) Correr un análisis de Covarianza

(2) Probar los Supuestos de:

(2.1) Independencia entre la Covariable y los tratamientos

(2.2) Regresión Lineal entre la Covariable X y la variable respuesta Y.

(2.3)Homogeneidad de las pendientes

(3) Obtener las medias mínimo-cuadráticas corregidas de los tratamientos

Para la solución del problema introducir los datos siguientes:

TRAT	Y	X
1	15	30
1	10	20
1	05	15
1	10	20
1	20	25
2	25	28
2	10	12
2	15	20
2	15	10
2	10	10
3	05	10
3	10	15
3	20	20
3	05	10
3	10	10

Ejemplo 28

Se realizó un experimento con frijol en cinco niveles de fertilización nitrogenada para medir el rendimiento en kilogramos por parcela (Y) con cuatro repeticiones (bloques), el cual fue afectado por una plaga de langostas (X).

Se registró el número de plantas afectadas por la plaga (X) y se ajustó por covarianza:

trat	0		0,5		1,0		1,5		2,0	
bloque	X	Y	X	Y	X	Y	X	Y	X	Y
1	10	3	14	2	11	04	11	3	3	1
2	09	4	18	3	16	02	10	1	8	2
3	08	2	11	1	05	01	09	3	7	1

4	13	3	20	3	07	02	07	2	5	4

Se pide:

(3) Correr un análisis de Covarianza

(4) Probar los Supuestos de:

(2.1) Independencia entre la Covariable y los tratamientos

(2.2) Regresión Lineal entre la Covariable X y la variable respuesta Y.

(2.3)Homogeneidad de las pendientes

(3) Obtener las medias mínimo-cuadráticas corregidas de los tratamientos

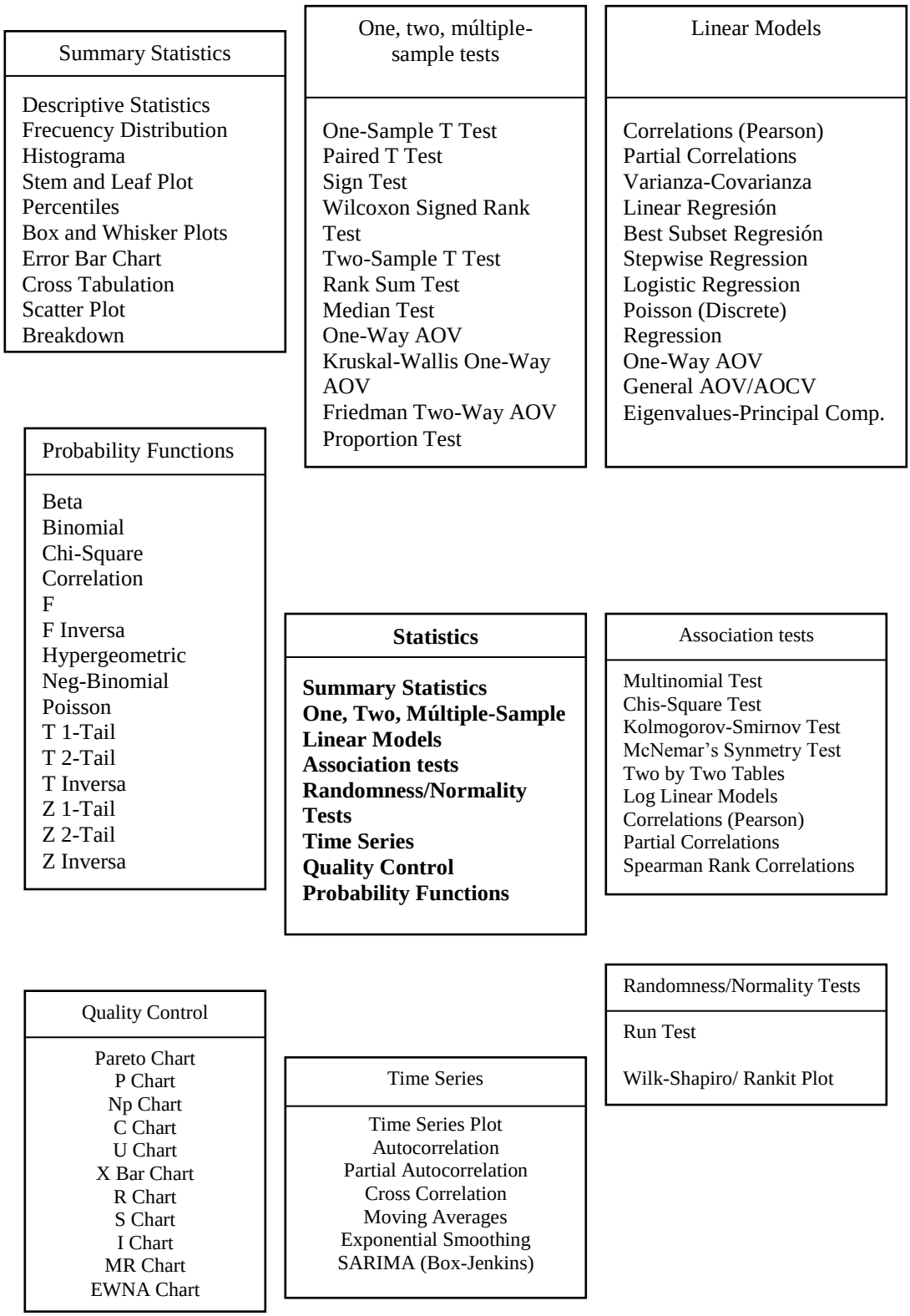
Bibliografía:

1. Berenson, M., Levine, D. (1996): *Estadística Básica en Administración*. PHH. México.
2. Bonilla, G. (1991): *Métodos Prácticos de Inferencia Estadística.* Trillas. México.
3. Daniel, W. (1988): *Estadística con aplicaciones a las Ciencias Sociales y a la Educación*. McGraw-Hill. México.
4. Mason, R., Lind, D., Marchall, W. (2000): *Estadística para la Administración y Economía.* Alfaomega. Colombia.

Statistix
for Windows

Programa interactivo
de Análisis Estadístico
para Computadoras Personales

Arbol de Ventanas del Statistix



Introducción al Paquete Estadístico STATISTIX

1. Es un programa estadístico muy versátil, fácil de aprender y usar. Se maneja enteramente por menús de pantallas; el usuario selecciona las actividades a realizar de un menú que está desplegado en pantalla.
2. Es un programa interactivo porque permite la comunicación entre usuario y software. Es un programa iterativo porque a una pregunta hay sólo una respuesta.
3. Fue desarrollado originalmente para la investigación en control estadístico de calidad e investigación de operaciones, pero por su gran facilidad de uso se popularizó en otros campos de la investigación científica.
4. Es un programa que cuenta con casi todas las técnicas estadísticas comúnmente empleadas tanto en la parte descriptiva, inferencial y diseños experimentales clásicos desarrollados en la tradición Fisheriana, así como herramientas para el control estadístico de la calidad **y** estudio de series de tiempo; incluso ha dado los primeros pasos en el análisis multivariante al incluir la técnica de componentes principales. Además posee técnicas para desarrollar análisis exploratorio de datos y análisis confirmatorio.
5. Este programa es más flexible que otros al permitir presentar gráficas para análisis, pensando en la publicación, para ello posee rutinas que facilitan enormemente que el usuario intitule las mismas, coloque el encabezado, rótulos de ejes abscisas y ordenadas y la fuente, con la opción "Title".
6. Es un software clasificado como WYSIWYG (acrónimo del término inglés What You See Is What You Get- lo que se ve es lo se obtiene),

esto significa que tal como se observa en pantalla así resulta al imprimir.

7. Existe en versiones para ambiente WINDOWS 98, 2000, Milenium. Los archivos generados en diferentes versiones son totalmente compatibles.

Su estructura se presenta en 7 opciones:

File, Edit, Data, STATISTICS (Menú Principal), Preferences, Windows, Help.

(ver en la página anterior el rnenú de opciones y el árbol de comandos del menú principal del STATISTIX)

El paquete introduce los datos en un spreadsheet u hoja. de trabajo, con un arreglo matricial dc 100 mil casos (registros u observaciones) y un máximo de 250 variables (versión 1). En la actualidad se conoce en el mercado la versión 7 (versión de prueba o Sherware) cuya vida útil es de un mes para trabajar con 8 variables, 10 mil casos y la versión 7, profesional, para 500 variables y 200 mil casos. Para adquirirlo visitar en Internet el sitio web del statistix en la dirección:

http://www.sigma-research.com/bookshelf/rtsxw.htm

Variables

Los tipos de variables que puede manejar el STATISTIX son muy diversas y pueden esquematizarse así:

(i) Variables numéricas:

Discretas tales como la edad en años cumplidos se escribe el nombre de la variable y a continuación se especifica colocando entre paréntesis la letra (I) de "interger".Este tipo de variable siempre se representa con cantidades enteras.

Continuas tales como el peso, la talla, presión sanguínea se escribe el nombre **y** después entre paréntesis se escribe la letra (R) de "real"; esta es la opción por defecto, de tal forma que si no se hace tal especificación de todos modos el paquete la asume.

Con este tipo de variables se representan todas las cantidades que pueden tomar un valor decimal, entero o fraccionario pero siempre dentro de un intervalo dado, de valores mínimo y máximo.

(ii) **Variables Alfanuméricas** **(**variables String) variables que representan una combinación de letras, números o símbolos. Consisten por lo general de variables cualitativas o nominales. El sexo, el estado civil, el tipo de profesión, los colores. En estos casos se. acostumbra escribir el nombre de la variable acompañado por una ese indicando que es una variable string o secuencia de caracteres seguida de un número que indica el ancho que ocupará el dato en cuestión, así: NOMBRE(**S25**), indica que la variable es un nombre y el ancho destinado para la variable es 25 caracteres. La "S" le viene de STRING, que significa que es **una secuencia de caracteres.**

Si se desea ingresar el sexo como "varón" **y** "hembra" se escribe SEXO(S6). en cambio, si se desea ingresarlo como **"masculino" y "femenino" escribiré SEXO(S9),** porque cada palabra posee un ancho máximo de 9 caracteres**.** Lo común es escribir el sexo codificado como cero y

uno**,** con el objetivo de poder utilizarlo dentro de un contexto de análisis de regresión, como una variable indicadora, ficticia o pseudovariable (variable dummy)

(iii) **Variables fechas:** son variables que registran la fecha de ocurrencia de un evento, como por ejemplo: en la administración intrahospitalaria interesa mucho estudiar la Fecha de Ingreso y la Fecha de Alta con las cuales se calcula la Tasa o Tiempo de Hospitalización. Se escribe el nombre de la variable y a continuación la**.** letra (D) de Date (Fecha). Por ejemplo, la fecha de nacimiento se escribe así: FECNAC(D)

Estas variables fechas son útiles en diversos campos de investigación, por citar algunos, en el campo de la reproducción animal nos encontramos con variables como intervalo entre partos, parto primer servicio, parto-concepción y otros.

Hasta aquí hemos enumerado las variables más comúnmente usadas, pero hay otras tales como:

(iv) **Variables de Asignación:** o creadas que son aquellas que. resultan de una combinación lineal de las variables existentes. En este caso, por ejemplo, si tenemos en la data las variables "altura" y "base", podríamos crear una nueva variable llamada "volumen", así:

VOL = PI * ALTURA * SQRT(BASE)

Es decir, una variable *"***VOL***"* que es el producto del valor PI = 3.1416, multiplicado por la altura y por la raíz cuadrada (SQRT) de la base.

O también, cuando creamos el índice de masa corporal (IMC) que es el resultado de dividir el peso sobre la talla al cuadrado como una expresión del grado de obesidad:

$$IMC = \frac{PESO}{TALLA^2}$$

En el Statistix en el menú de transformaciones puede ser creada escribiéndola así:
IMC = PESO / TALLA^2

O también cuando queremos realizar un ajuste del peso a los 205 días (en ganado de carne):

PA205 = ((PF-Pl) / EDAD)*205 + PI

O calcular la ganancia diaria como:

GANDIA = (PF-PI) / EDAD

TIPOS DE DATOS:
Con el Statistix pueden utilizarse:

(i) Datos numéricos: números reales, números en notación científica, por ejemplo:

$3.44x10^{-3}$ se escribirá como 3.44 **E(-3)**

(ii) **Nombres:** se escriben nombre que no posean más de 8 letras.

(iii) **Datos perdidos o faltantes:** se escriben con "M" de missing. Esta función la realiza el Statistix automáticamente.

(iv) **Datos binarios o dicotómicos:** de 0,1 para variables indicadoras o variables dummy.

En relación a la **creación de nuevas variables** es importante señalar que una estrategia muy utilizada cuando los datos no cumplen los supuestos de normalidad y homogeneidad de varianzas es recurrir a la transformación de la data, utilizando para ello la opción **"Transformations"**del paso "data". Entre.

estas transformaciones podemos mencionar sólo algunas de las 59 disponibles en el STATISTIX.

Si queremos obtener la raiz de X y asignarsela a una nueva variable (RAIZX), escribiremos así:

RAIZX = SQRT(X)

Si queremos elevar a una potencia, por ejemplo, X al cuadrado:

X2=X^2

También utilizando el menú de transformaciones podemos elevar al cuadrado:
X2 = SQR(X)

Si queremos elevar a la quinta potencia, escribiremos así:

X5 = X ^5

Para obtener el logaritmo:

de base 10 : **LX** = **LOG(X)**

de base **2.72: LX = LN(X)**

Si queremos las potencias del número "e":

Por ejemplo: "e" elevado a la menos 8, se escribirá así:

A = EXP(-8)

Si **queremos crear variables normalizadas con media cero y varianza uno, escribiremos (ESTANDARIZACION DE VARIABLES):**

NORMALX = NO(X)

Una suma acumulada de X se produce así:

ACUMX = **CU(X)**

Las **Funciones Trigonométricas** se calculan así:

SENOX = SIN(X)

COSX = COS(X)

Operadores Aritméticos:

A = B + C (variable de suma)

A = B - C (variable de resta)

A = B * C (variable producto)

A = B / C (variable cociente)

ACUBO = **A^3** (variable exponenciación)

Expresiones lógicas:

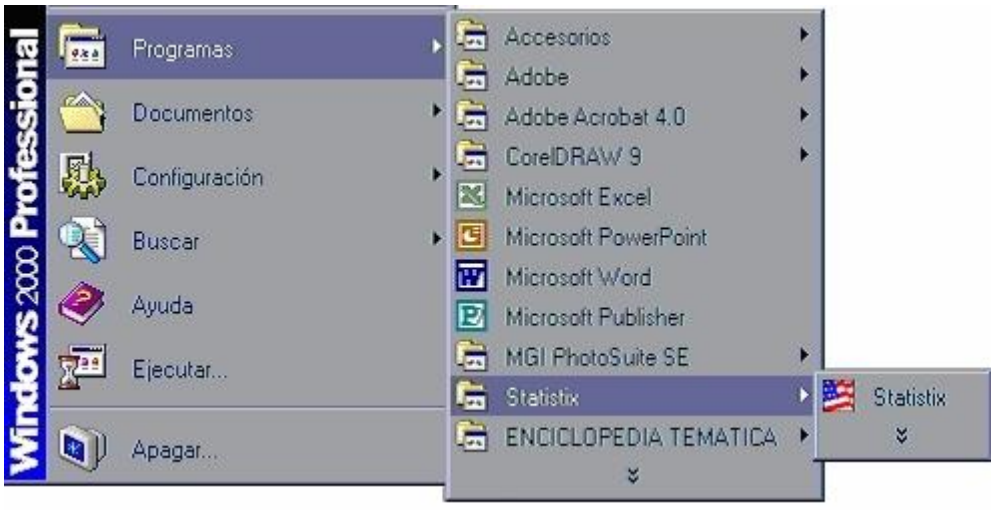
Operadores de comparación:

Mayor que	>
Menor que	<
Diferente	<>
Menor o igual	<=
Mayor o igual	>=

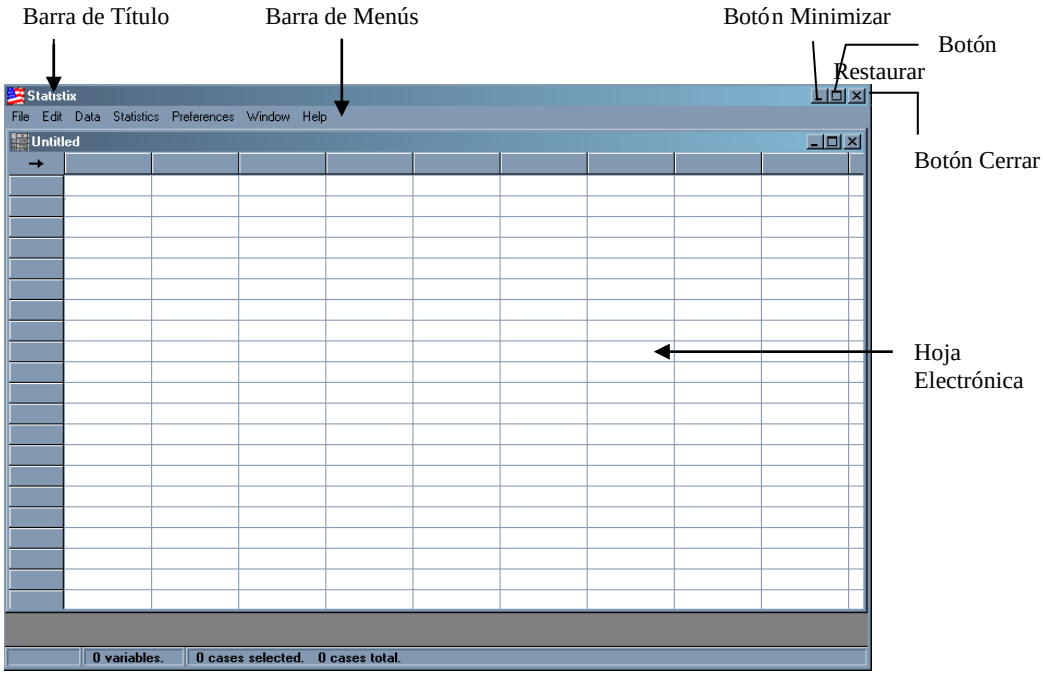
Abrir Statistix

Para iniciar una sesión de trabajo con Statistix realice lo siguiente.

⇨ Haga clic en el botón **Inicio** de la Barra de tareas de Windows, seleccione **Programas** y escoja Statistix.



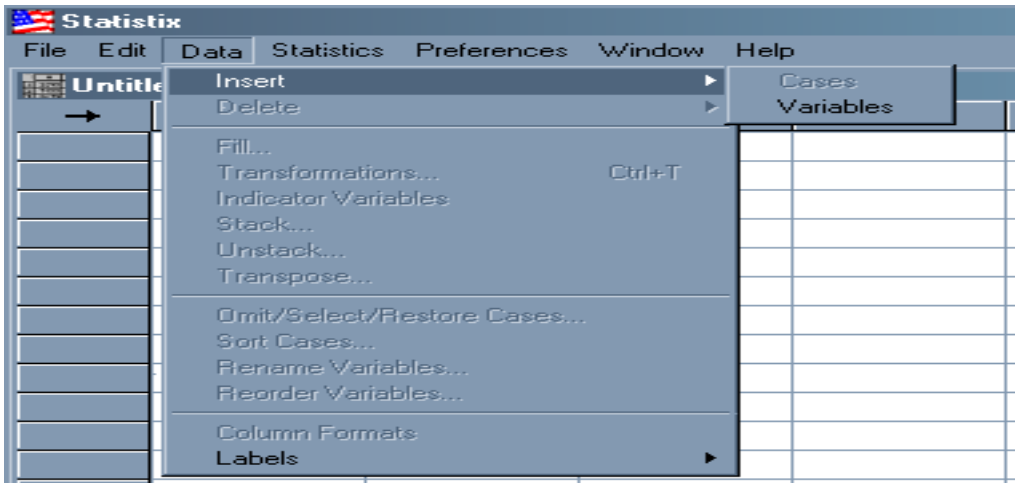
⇨ Aparecerá la ventana de trabajo de Statistix.



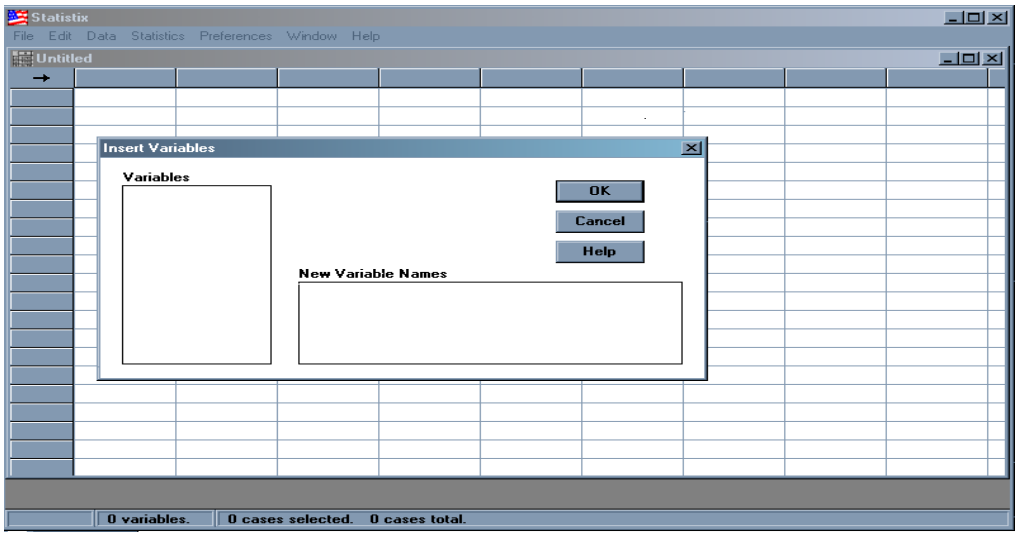
Para crear una base de datos

Pasos:

1. En la Barra de Menús selecciono la opción **DATA** a continuación **INSERT y** seguidamente **VARIABLES:**

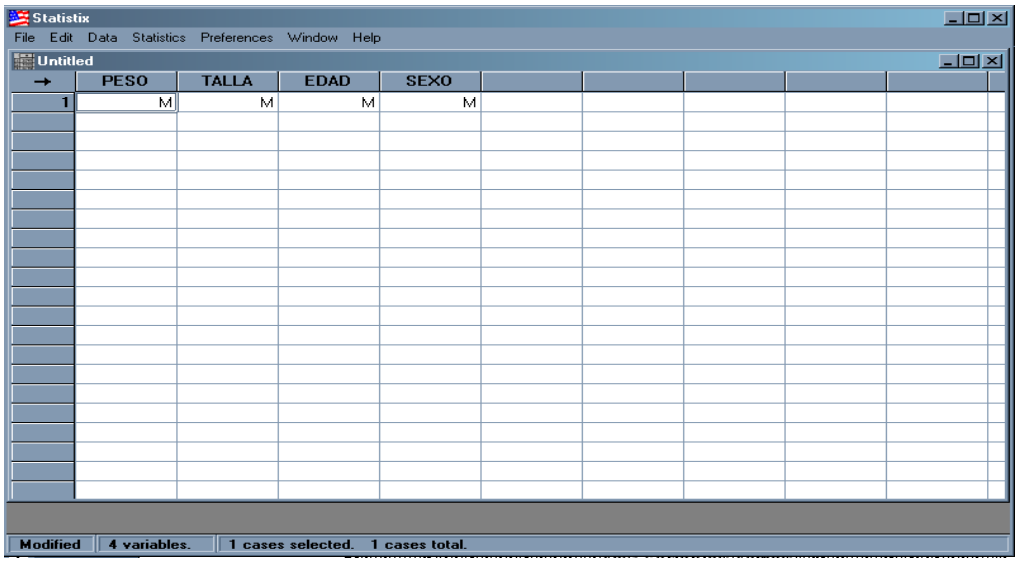


Al pulsar VARIABLES aparecerá el cuadro de diálogo que sigue:



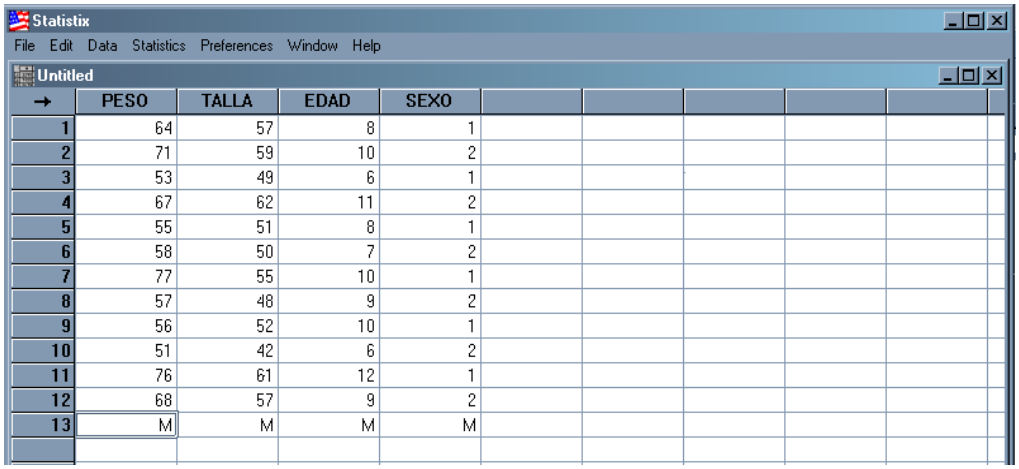
En el cuadro donde donde dice **New Variables Names** se indica el nombre de las nuevas variables.

Si estas variables son: PESO, TALLA, EDAD y SEXO, aparecerá tal como se indica a continuación:



A continuación comenzamos a introducir los datos de peso, talla, edad y sexo, presionando ENTER después de ingresar cada dato.

La hoja electrónica ahora tendrá el aspecto siguiente:



TABULACION CRUZADA

Hágase una clasificación bidireccional (por sexo y afiliación política) de los miembros de la Unión Política Rodees (R = republicano, D = demócrata, I = independiente, V = varón, M = mujer):

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
R	X					X		X									X					X		X							X	
D									X		X				X				X	X	X					X		X		X		
I		X	X	X	X					X		X	X	X		X		X						X		X		X				X
V	X	X			X	X	X							X	X		X	X		X						X		X	X	X		
M			X	X				X			X	X	X			X			X		X	X	X	X		X				X	X	X

Para realizar la tabulación cruzada se procede así:

1. **Inicio**
2. **Programas**
3. **Statistix**
4. **Data**
5. **Insert**
6. **Variables**
7. **AFILIA SEXO**
8. **Introduzco datos:**
   100 1
   001 1
   001 2
   001 2
   .
   .
   .
   001 2
9. **File**
10. **Save**
11. **UNIONPOL**
12. **STATISTICS**
13. **SUMMARY STATISTICS**
14. **CROSS TABULATION**
15. **VER TABLA CRUZADA E INTERPRETAR**

El programa Statistix presenta la siguiente salida:

STATISTIX FOR WINDOWS

CROSS TABULATION OF AFILIA BY SEXO

	VARON	MUJER	
AFILIA	**1**	**2**	**TOTAL**
001 (Republicano)	**9**	**7**	**16**
010 (Demócrata)	**4**	**5**	**9**
100 (Independiente)	**3**	**4**	**7**
TOTAL	**16**	**16**	**32**

DISTRIBUCIÓN DE FRECUENCIAS PARA DATOS AGRUPADOS

Cuando se tiene una gran cantidad de números, el proceso de extraer conclusiones sobre los mismos se hace lento y tedioso. Por esta razón, los datos son "comprimidos" de forma tal que su análisis se facilite.

Observa los datos que se dan a continuación. Los mismos corresponden a las notas obtenidas por 100 alumnos del curso de inglés de 7mo grado en el liceo Cruz Salmerón Acosta de Cumaná.

6	20	8	19	9	17	11	16	10	15
12	14	13	15	14	16	14	16	13	17
16	13	16	6	13	14	12	12	17	15
10	16	12	16	12	6	14	15	13	14
13	12	15	14	10	13	12	16	10	16
17	13	14	8	13	11	16	13	15	12
12	13	10	13	18	14	13	9	17	18
14	15	14	12	14	12	14	18	12.	14
13	12	19	13	15	20	12	13	19	17
11	15	14	12	13	14	14	14	15	12..

Notas obtenidas por los alumnos del 7 mo.grado de Inglés del Liceo Cruz Salmerón Acosta de Cumaná.

.STATISTIX FOR WINDOWS

FREQUENCY DISTRIBUTION OF NOTA

LOW	HIGH	FREQ	PERCENT	CUMULATIVE FREQ	CUMULATIVE PERCENT
6.0	**7.7**	**3**	**3.0**	**3**	**3.0**
7.7	**9.4**	**4**	**4.0**	**7**	**7.0**
9.4.	**11.1**	**8**	**8.0**	**15**	**15.0**
11.1	**12.8**	**16**	**16.0**	**31**	**31.0**
12.8	**14.5**	**35**	**35.0**	**66**	**66.0**
14 .5	**16.2**	**20**	**20.0**	**86**	**86.0**
16 .2	**17.9**	**6**	**6.0**	**92**	**92.0**
17 .9	**19.6**	**6**	**6.0**	**98**	**98.0**
19.6	**21.3**	**2**	**2.0**	**100**	**100.0**
TOTAL		**100**	**100.0**	**-**	**-**

PARA ORGANIZAR LOS DATOS EN UNA DISIRIBUCION DE FRECUENCIAS SE PROCEDE ASI:
PASOS:

1. **BUSCAR EL RANGO = MAYOR – MENOR (RANGO = 20-6 =14)**
2. **BUSCAR EL INTERVALO DE CLASE (i):**

REGLA DE STURGES: = RANGO / NUMERO DE CLASES =
14 / 1+3.32 LOG(n) = 14 / 7.64 = 1.8

REGLA EMPIRICA: RANGO / NUMERO DE CLASES =

$14/\sqrt{n} = 14/10 = 1.4$

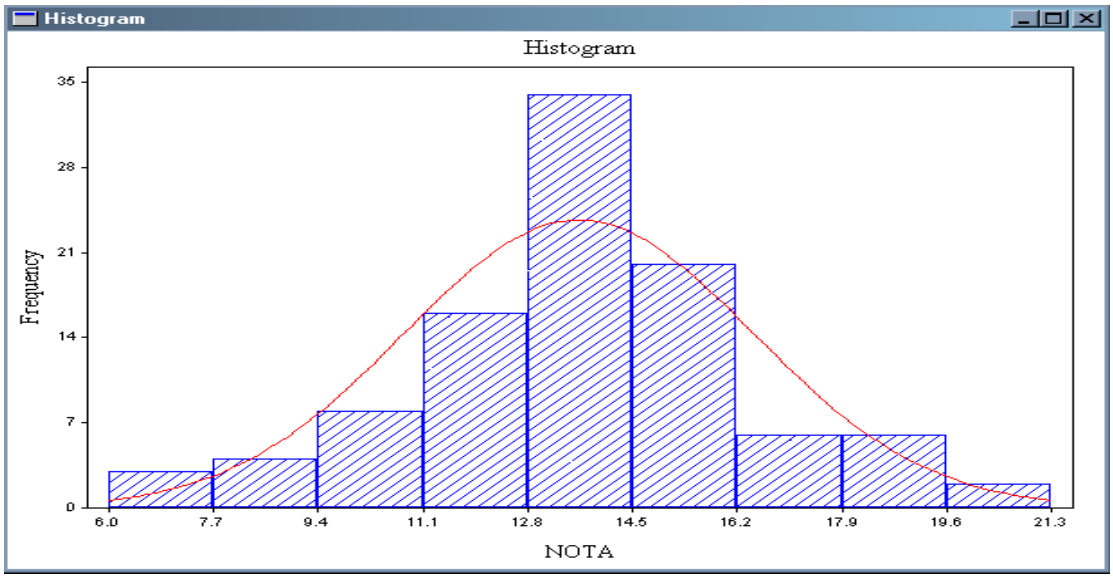
REGLA DE YULE: RANGO / NUMERO DE CLASES =

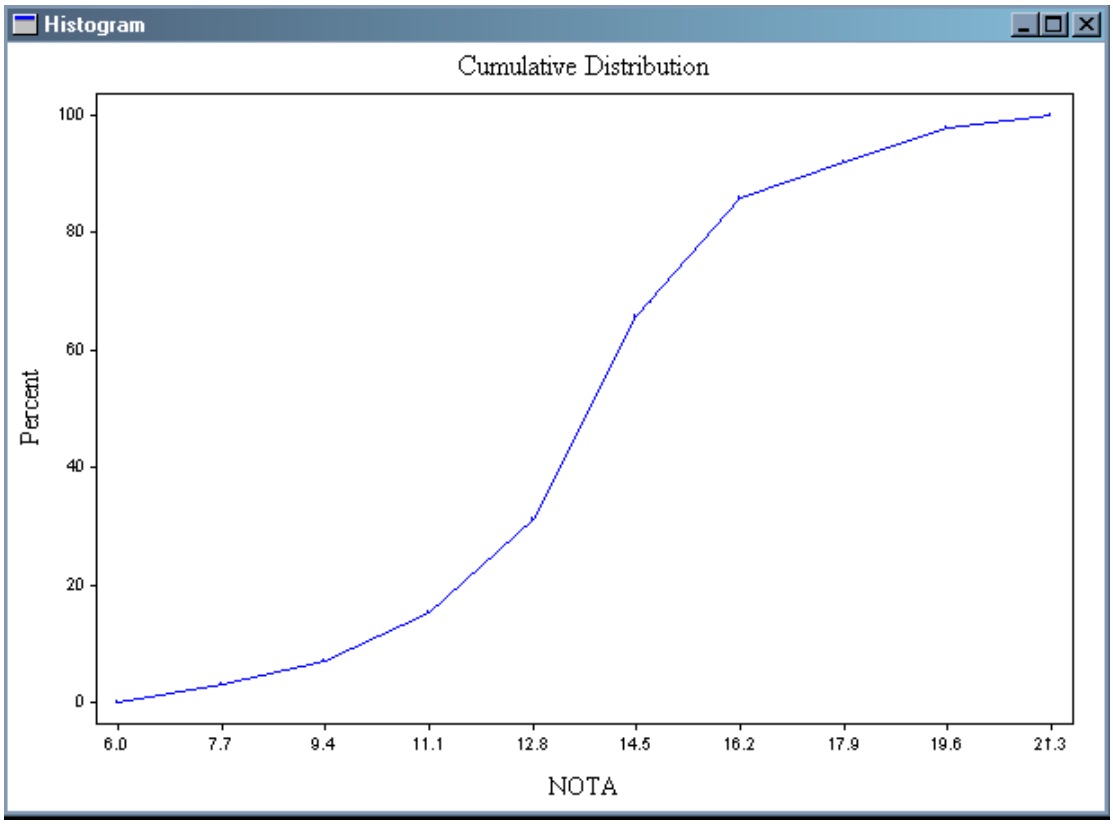
$14/2.5\sqrt[4]{n} = 14/1.9 = 1.7$

Se decide según la regla de Yulé utilizar un i =1.7 e iniciar las clases en seis (6)
El paso uno y dos son ejecutados manualmente, es decir, con una calculadora.

(3) Secuencia a seguir en el software:

1. **INICIO**
2. **PROGRAMAS**
3. **STATISTIX**
4. **DATA**
5. **INSERT → VARIABLES**
6. **NOTA (Nombre de la Variable)**
7. **INTRODUCIR DATOS:**
8. **6, 12, 16, 10, 13, ... , 12**
9. **FILE**
10. **SAVE**
11. **LICEO**
12. **STATISTICS**
13. **SUMMARY STATISTICS**
14. **FRECUENCY DISTRIBUTION**
15. **NOTA**
    a. **6**
    b. **20**
    c. **1.7**
    d. **ok**


Histogram
Histogram
Frequency
35
28
21
14
7
0
6.0
7.7
9.4
11.1
12.8
14.5
16.2
17.9
19.6
21.3
NOTA


Histogram
Cumulative Distribution
Percent
100
80
60
40
20
0
6.0
7.7
9.4
11.1
12.8
14.5
16.2
17.9
19.6
21.3
NOTA

OPERADORES BOOLEANOS:

AND (cumplir con ambas condiciones de la sentencia)
OR (cumplir con una de las dos condiciones de la sentencia)
NOT (no cumplir la condición de la sentencia)

EXPRESIONES CONDICIONALES DEL TIPO IF-THEN-ELSE:

IF VARIEDAD = 'L6014' THEN TRAT = *14*
lF TRAT = 13 THEN DELETE
lF ZAFRA = 1 THEN CORTE = *'P'*
IF 7 <= PESO < = 10 THEN MASA = *A* ELSE MASA = B

TIPOS DE ARCHIVOS:

ASCII (XXXX .TXT, XXXX . PRN, XXXX . DAT)

DBASE (XXXX . DBF)

EXCEL (XXXX . XLS) (versiones 2 al 4)

LOTUS (XXXX . WK1, XXXX . WK2, XXXX . WK4)

QUATTRO PRO (XXXX . WQ1, XXXX. WB1)

STATISTIX XXXX . SX (archivos de DOS y Windows)

DISEÑOS EXPERIMENTALES

ESPECIFICACION DEL MODELO ADITIVO LINEAL:

Diez observaciones son importantes en la especificación del modelo aditivo lineal:

(i) Se tiene que incluir el término o factor relacionado con repeticiones o réplicas como una variable. Se acostumbra a denominar REP.

(ii) Se acostumbra declarar dentro del modelo el término relacionado con el error, aunque cuando es un caso de diseño sencillo, lo asume por defecto.

(iii) No se deben declarar más de 50 niveles por factor.

(iv) No se deben declarar más de 7 covariables.

(v) No se deben incluir más de 50 variables independientes en un análisis de regresión.

(vi) No se pueden incluir más de 50 valores faltantes o perdidos en un mismo ensayo.

(vii) El número de repeticiones por combinación de niveles de tratamientos en un factorial no puede ser superior a 50.

(viii) No se puede trabajar con más de 250 variables.

(ix) No se puede trabajar con más de 100 mil casos (registros u observaciones)

(x) No pueden declarar más de tres términos de error en un mismo ensayo.

PRINCIPALES MODELOS DE DISEÑOS EXPERIMENTALES:

El diseño completamente al azar se especifica así:

Y = TRAT REP * TRAT(E)

o también: Y = TRAT

El diseño en bloques al azar:

Y = TRAT BLOQ BLOQ * TRAT(E)

o también: Y = TRAT BLOO

El diseño factorial completo de 3 factores:

Y = A B C A*B A*C B*C A*B*C REP*A*B*C(E)

o también: Y ALL(X1 X2 X3)

El diseño completamente anidado de 4 factores:

Y = Xl X1*X2 X1*X2*X3 X1*X2*X3*X4 REP*X1*X2*X3*X4(E)

El diseño de parcelas divididas: con una parcela principal Xl, con una subparcela X3 y arreglo en bloques X2:

Y = Xl X2 X1*X2(E) X3 X2*X3(E) Xl*X3 X1*X2*X3(E)

El diseño de mediciones repetidas (tratamiento, sujeto y repetición):

Y = TRAT SUJ TRAT*SUJ*REP

DISEÑOS EXPERIMENTALES CLASICOS.

1. ANOVA PARA UN DISEÑO COMPLETAMENTE AL AZAR:

Se desea verificar si el tiempo de llenado (en segundos) de cereales en cajas de 368 gramos difiere según el tipo de máquina (1, 2, 3):

MAQUINAS		
1	2	3
25.40	**23.40**	**20.00**
26.31	**21.80**	**22.20**
24.10	**23.50**	**19.15**
23.74	**22.75**	**20.60**
25.10	**21.60**	**20.40**

PASOS:

(1) INICIO
(2) PROGRAMAS
(3) STATISTIX
(4) DATA
(5) INSERT (CASES, VARIABLES) ⟶ VARIABLES
(6) CREACION DE LAS VARIABLES
(7) INTRODUCIR DATOS:

(8) MAQ	Y	REP	OK
	1	25.40	1
	1	26.31	2
	1	24.10	3
	1	23.74	4
	1	25.10	5
	2	23.40	1
	2	21.80	2
	.	.	.
	.	.	.
	.	.	.
	3	20.40	5

(9) SAVE CEREAL OK
(9) LINEAR MODELS
(10) GENERAL AOV /AOCV
(11) INDICAR CUAL ES LA VARIABLE DEPENDIENTE: Y
INDICAR EL MODELO: MAQ REP *MAQ(E)
(12) VER EL ANOVA E INTERPRETAR

2. DISENO DE BLOQUES AL AZAR

Se desea evaluar el efecto que tienen tres zonas de ventas (norte, sur y oeste) y tres precios (P1=129, P2=139, P3=149) sobre el volumen de ventas (número de unidades vendidas) de un producto:

	bloques (zonas)		
trat	norte	sur	oeste
Pl	915	909	890
P2	935	880	860
P3	920	920	880

Realizar un ANOVA y concluir.
PASOS:

(1) INICIO
(2) PROGRAMAS
(3) STATISTIX
(4) DATA
(5) INSERT VARIABLES

TRAT BLOQ REP Y
INTRODUCIR DATOS:

1	1	1	915
2	1	2	935
3	1	3	929
1	2	1	909
2	2	2	889
3	2	3	920
1	3	1	890
2	3	2	860
3	3	3	885

SAVE VENTAS

LINEAR MODELS

GENERAL AOV /AOCV
VD = Y

MODELO: TRAT BLOQ BLOQ * TRAT(E)
VER ANOVA E INTERPRETAR

3. EXPERIMENTO FACTORIAL CRUZADO

Para realizar una auditoria, se desea investigar el efecto de utilizar dos diferentes métodos Al y A2, y dos diferentes tiempos de proceso de la auditoria B1 y B2, midiendo como variable de respuesta el porcentaje de eficiencia:

Métodos de Auditoria	B1 (Tiempo 2 horas)	B2 (Tiempo 4 horas)
A1	70	80
	68	76
	74	72
A2	76	70
	68	82
	64	68

PASOS:

(1) INICIO
(2) PROGRAMAS
(3) STATISTIX
(4) DATA
(5) INSERT VARIABLES
(5) METODO TIEMPO CELDA REP Y
(6) MATRIZ DATOS:

1	1	11	1	70
1	1	1	2	68
1	1	1	3	74
2	1	2	1	76
2	1	2	2	68
2	1	2	3	64
1	2	3	1	80
1	2	3	2	76
1	2	3	3	72
2	2	4	1	79
2	2	4	2	82
2	2	4	3	68

(8) SAVE ⟶ AUDITOR
(9) LINEAR MODELS
(l0) GENERAL AOWAOCV
(11) VD: Y
MODELO: MEJODO TIEMPO METODO*TIEMPO REP*METODO*TIEMPO(E)
(12) VER ANOVA E INTERPRETAR

4. ANALISIS DE REGRESION MULT1PLE:

Se desea encontrar la ecuación que permite relacionar el peso en función de la talla y edad.

PESO	(Y)	64	71	53	67	55	58	77	57	56	51	76	68
TALLA	(X1)	57	59	49	62	51	50	55	48	52	42	61	57
EDAD	(X2)	8	10	6	11	8	7	10	9	lO	6	12	9

PASOS:

(1) INICIO
(2) PROGRAMAS
(3) STATISIIX
(4) DATA
(5) INSERT ⟶ VARIABLES
(6) Y X1 X2

(7) MATRIZ DE DATOS:

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

(8) SAVE ⟶ ESCOLAR

(9) LINEAR MODELS

(10) LINEAR REGRESSION

(11) VA = Y

MODELO: Xl X2

(12) VER ECUACION DE REGRESION E INTERPRETAR COEFICIENTES

Bibliografía

1. ARMITAGE,P.;BERRY,G.: Estadística para la Investigación Biomédica. DOYMA. 1978. p.656.

2. AMERICAN SOCIETY OF ANIMAL SCIENCE: Techniques and procedures in production research, New York, A.S.A.S., 1963, p.228.

3. ANDERSON, R. L., and BANCROFT, T. A.: Statistical theory in research, New York, McGraw Hill, 1962, p. 464.

4. ANDERSON, V. L., and McCLEAN, R. A.: Design of experiments, New York, Marcel Dekker, 1974, p. 418.

5. ARNAU, J. A.: Diseños Experimentales en Psicología y Educación, México, Trillas, 1986, p. 406.

6. BARRY, A., y OTROS : Curso de Estadística Experimental Avanzado, Lima (Perú), Ministerio de alimentación, 1979, p. 444.

7. BERENBLUT, I. I.: Change-Over Design with Complete Balance for first residual effects, Biometrics, 1964, 44: pp. 707- 712.

8. BOX, G.; HUNTER, W., and HUNTER, J.: Statistic for experimenters (An introduction to design, data analysis and model building), New York, John Wiley & Sons, 1978, p. 563.

9. BRANDT, A. E.: Test of significance in reversal o switchback triáis, Iowa, Agr. Expt. Sta, Research Bull, 1938, p. 234.

10. BROWN, B. W.: The Cross-Over Experiment for Clinical Triáis, Biometrics, 1980, 36: pp. 69-80.

11. CASTEJÓN, O.: Introducción al Diseño y Análisis de Experimentos. Mimeografiado. Trabajo de Ascenso.LUZ. Vol I. 1985.p.200.

12. CASTEJÓN, O.: Introducción al Diseño y Análisis de Experimentos. Mimeografiado. Trabajo de Ascenso.LUZ. Vol II. 1989.p.200.

13. CASTEJÓN; O.: Modelos de Regresión Curvilineales. Mimeografido. Trabajo de Ascenso. 1990. LUZ

14. CALZADA BENZA, J.: Métodos estadísticos para la investigación, 3ra. ed., Lima (Perú), Jurídica S.A., 1970, p. 643.

15. CANTATORE DE FRANK, N.: Manual de estadística aplicada, 2da. ed., Buenos Aires (Argentina), Hemisferio Sur S.A., 1980, p. 395.

16. ------_ .; PEREYRA DE GALLO, LL.: Experimentos con animales, Buenos Aires (Argentina), Facultad de Ciencias Veterinarias, Universi_ dad de Buenos Aires, 1978, p. 160.

17. Chacín, F.: Diseño y Análisis de Experimentos. Ediciones Vicerrectorado Académico de UCV. 2000. p.387

18. COCHRAN, W. G., y COX, G. M.: Diseños Experimentales, México, Trillas, 1976, p. 661.

19. .; AUTREY, K., and CANNON, C.: A Double Change-Over Design for Dairy Cattle Feeding Experiments, J. Dairy Sci, 1941, 24: pp. 937- 951.

20. DANIEL,W.: Bioestadística. Base para el Análisis de las Ciencias de la Salud. Limusa-Wiley. 2002.p. 148.

21. DAS, M. N., and GIRI, N. C: Design and analysis of experiments, New York, John Wiley & Sons, 1979, p. 295.

22. EBBUTT, A. F.: Three-Period Crossover Design for two Treatments, Biometrics, 1984, 40: pp. 219-224.

23. FEDERER, W. T.: Experimental design (Theory and application), New York, The Mac.Millan Co., 1955, p. 591.

24. GILL, J. L.: Design and Analysis of Experiments in the Animal and Medical Sciences, lowa, The lowa State University Press, 1981, Vol. II, pp. 169- 260.

25. _____ .: Current status of múltiples comparisons of means in designed experiments, J. Dairy Sci., 1973, 56: p. 973.

26. GRIZZLE, J. E.: The two period change-over design and its use in clinical triáis, Biometrics, 1965, 21: pp. 467- 480.

27. GUTIÉRREZ; H.;Vara, R.: Análisis y Diseño de Experimentos. McGraw_Hill. 2004. p.571.

28. HICKS, C: Fundamental concepts in the design of experiments, New York, Rinehart and Winston, 1973, p. 430.

29. HILLS, M., and ARMITAGE, P.: The two-period crossover clinical trial, Bristish, J. of Clinical Pharmacology, 1979, 8: pp. 7- 20.

30. JEROME, L. : Introduction to statistical inference, Michigan, Brothers, Ann Arbor, 1957, p. 568

31. JOHN, W. M.: Statistical design and analysis of experiments, New York, The MacMillan, 1971, p. 356.

32. KENDALL, M. G., and BUCKLAND, W. R.: Diccionario de estadística, Madrid (España), Pirámide, 1976, p. 384.

33. KUEHL, R. : Diseños de Experimentos. Principios Estadísticos para el diseño y análisis de investigaciones. Thomson Editores. 2001.p.666.

34. LASKA, E.; MEISNER, M., and KWSHNER, H. B.: Optimal Crossover Design in the Presence of Carryover Effects,Biometrics, 1983,39:pp.1087-1091.

35. LI, C. C.: Introducción a la estadística experimental, 2da. ed., Barcelona (España), Omega, 1969, p. 496.

36. LISON, L.: Estadística aplicada a la biología experimental, Buenos Aires (Argentina), Eudeba, 1976, p. 357.

37. LITTLE, T. M. y HILL, F. J.: Métodos estadísticos para la investigación en la agricultura, México, Trillas S.A.,1976, p. 270.

38. LUCAS, H. L.: Design and analysis of feeding experiments with milking dairy cattle, North Carolina, Institute of Statistics Mimeo Series, n? 18,1974,p.473.

39. _ .: Switch-back triáis for more than two treatments, J. Dairy Sci., 1956, 39: pp. 146- 154.

40. _ .: Extra-period Latin Square Changeover design, J. of Dairy Sci., 1957, 4: pp. 225- 239.

41. MÉNDEZ, I.: Modelos estadísticos lineales, México, Fondo de Cien cia y Cultura Audiovisual, 1976, p. 140.

42. MENDENHALL, W.: Introducción a la probabilidad y estadística, New York, Wadsworth Internacional, 1982, p. 626.

43. _ .: The Design and Analysis of Experiments,. California, Duxbury Press, 1968, p. 465.

44. Martínez, A.: Diseños Experimentales. Métodos y Elementos de Teoría. Trillas.1988. p. 756.

45. MILTON, S.: Estadística para Biología y Ciencias de la Salud. McGraw-Hill-Interamericana.2001. p.592.

46. MYERS, R. H.: Response surface methodology, Boston, Allyn and Bacon, 1971, p. 388.

47. MONTGOMERY, C.: Design and Analysis .of experiments, New York, John Wiley, 2004., p. 686.

48. MONZÓN, D.: Introducción al diseño de experimentos, Maracay, (Venezuela), Ministerio de Agricultura y Cría, Centro de Investigaciones Agronómicas, 1964, p. 167.

49. _ .: Sobre la escogencia de un denominador apropiado de la prueba de razón de dos varianzas, MAC, Centro de Investigaciones Agropecuarias, Maracay, Circular Nº 8, 1963, pp. 3- 38.

50. OGAWA, J.: Statistical theory of the analysis of experimental design, New York, Marcel Dekker, 1974, p. 653.

51. OSTLE, B.: Estadística aplicada, México, Limusa Wiley, 1965, o. p. 629.

52. PAEZ, B.: Métodos de investigación en producción animal, Costa Rica, Centro Tropical de Enseñanza e Investigación, Instituto de Ciencias Agrícolas, 1965, p. 267.

53. PANSE, V.,SUKHATME: Métodos Estadísticos para Investigadores Agrícolas. Fondo de Cultura Económica.1959. p.347.

54. PATTERSON, H. D.: The Construction of Balanced Design for Experi_ ments involving sequences of treatments, Biometrics, 1959, 39: pp. 32- 48.

55. _ ., and LUCAS, H. L.: Change-Over designs, Tech. Bull,

147, N. Carol.,agri. Exp. Stn., 1962, pp. 110- 129.

56. _ .: Extra-Period Change-Over Design, Biometrics, 1959,15: pp. 116-132.

57. PIMENTEL, F.: Curso de estadística experimental, Argentina, Hemisferio Sur, 1978, p. 323. .

58. REYES, P.: Diseño de experimentos aplicados, México, Trillas, 1978, p. 344.

59. RODRÍGUEZ,J.: Métodos de Investigación Pecuaria. Trillas. 1991.p.208.

60. SEATH, D. M. : 2x2 Factorial design for double-reversal feeding experiments, J. Dairy Sci., 1944, 27: p. 159.

61. SOKAL, R. R. y ROHLF, F. J.: Biometría, Madrid, H. Blume Ediciones, 1979, p. 832.

62. SNEDECOR, G. W., y COCHRAN, W. G.: Métodos estadísticos, Iowa, The Iowa State University Press Ames, 1967, p. 703.

63. STEEL, R. G., and TORRIE, B. H.: Principies and procedures of statistics with special reference of the biological science, New York, McGraw Hill Book Co., 1930, p. 633.

64. STOBBS, T. H., and SANDLAND, R. L.: The use of latin square change-over designs with dairy cows to detect differences in quality of tropical pastures, Aust. J. exp. Agrie. Anim. Husb., 1972, 12: pp. 463- 469.

65. TAYLOR, W., and ARMSTRONG, P.: The efficieney of some experimental designs used dairy husbandry experiments, J. Agri. Sci., 1953, pp. 43-407.

66. VILLASMIL, J. J.: Apuntes de clase del curso de diseños experi mentales ofrecido en el postgrado de producción animal, División de Postgrado de Agronomía y Veterinaria, L.U.Z., 1983, p. |60|.

67. WALPOLE, R., and MYERS, R.: Probabilidad y estadística para ingenieros, México, Interamericana S.A., 1982, p. 578.

68. WALDO, D.: An evaluation of múltiple comparison procedures, J. of animal science, 42 (N2 2), 1976, pp. 34- 39.

69. WINER, B.: Statistical principies in experimental design, 2nd. ed., New York, McGraw Hill Book Co., 1962, p. 672.

70. WILLIAMS, E. J.: Experimental Designs Balanced for the Estimation of residual effects of treatments, Australian J. Sci.Res.A, 1949, 2: pp. 149- 168.

Webgrafía

1. http://www.monografias.com/trabajos/cuasi/cuasi.shtml

2. http://www.udc.es/dep/mate/estadistica2/sec2_6.html

3. http://math.uprm.ed/edgar/miniman10.ppt#256,1,10.DISENOS EXPERIMENTALES

4. http://math.uprm.ed/edgar/miniman10.ppt#270,15,ejemplo de anova

5. http://www.efdeportes.com/edf46/invest.htm

6. http://html.rincondelvago.com/disenos-experimentales-de-investigacion.html

7. http://agronomia.unal.edu.co/documentos/ejercici.pdf

8. http://www.conocimientos.net/dcmt/ficha1147.html

9. http://www.eumed.net/libros/2006c/203/2f.htm

10. http://www.matematica.ues.edu.sv/trabajosdegraduacion/desarrollo/Apendice.pdf

11. http://www.eumed.net/tesis/amc/52.htm

12. http://www.udc.es/dep/mate/estadistica2/sec2_3.html