Optimiza 12

LIBRO 5
Diseño de experimentos

Ing. Alejandro Roberti

Ing. Gustavo Chijani — Ing. Verónica Esain — Ing. Esteban Gidekel

Capítulo 19. Diseño de Experimentos (1° Parte)

Un experimento diseñado, o un diseño de experimentos, se define como un programa que contempla la realización de una prueba (o serie ordenada de pruebas) en las que se introducen cambios deliberados en las variables de entrada de un proceso o sistema, de manera que sea posible observar o identificar las causas de los cambios en la respuesta de salida.

Noticia histórica

La técnica del Diseño de Experimentos apareció a inicios del siglo XX, pero, obviamente, la experimentación es antigua como el ser humano. Durante muchos años se utilizó la estrategia *ceteris paribus*, o de un-factor-por-vez, reconocida como un método científico y atribuida a Francis Bacon en el siglo XVII, que en realidad derivaba de los griegos. La estrategia de "un factor-por-vez" consiste en ir modificando cada vez un solo factor (variable) y así evaluar los efectos que cada uno de ellos por separado tiene sobre la respuesta. Esta estrategia, a pesar de ser utilizada en empresas y seguir el "método científico", es ineficiente — como veremos más adelante — para hallar las mejores condiciones del proceso.

Esa metodología tuvo su apogeo con Thomas Edison, quien la aplicó para alguno de sus inventos y quedó obsoleta cuando Ronald Fisher[1], en la década de 1920 desarrolló un método más eficiente para experimentar basado en los diseños factoriales.

El diseño de experimentos fue aplicado por primera vez por Fisher en Inglaterra en la agricultura, y sus experiencias fueron publicadas en 1935 en su libro "*Design of Experiments*". Sus estudios estaban centrados en mejorar la producción de papas, trabajando para la Estación Agrícola Experimental de Rothamstede en Londres[2].

Los Diseños Experimentales fueron utilizados por los agrónomos, (aproximadamente desde 1929), como un medio idóneo para planificar la

---

[1] Ronald Aylmer Fisher (1890 – 1962) Estadístico y biólogo que usó la matemática para combinar las leyes de Mendel con la selección natural, de manera que ayudó así a crear una nueva visión conocida como síntesis evolutiva moderna. Trabajó en una estación agrícola experimental (Rothamsted Research), donde desarrolló el análisis de la varianza para los grandes datos de sus cultivos desde 1840. Allí se afianzó como bioestadístico.

[2] Citado por Víctor Salas Aranda, "Diseño de experimentos con software estadístico". Universidad de Sevilla, junio 2018.

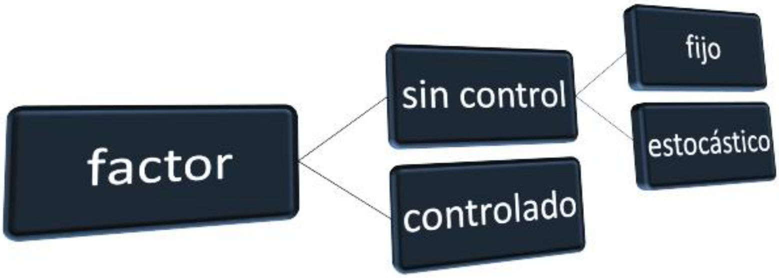
realización de ensayos realizados modificando simultáneamente numerosas variables y alternativas como, por ejemplo, la evaluación simultánea de distintas variedades de semillas y de diferentes tratamientos. Esta metodología — es más apropiado llamarla “conjunto de métodos” — tuvo una difusión lenta, por un lado, debido a su aspecto renovador y por otro debido a la necesidad de realizar, en algunos casos, un importante número de cálculos.

Desde hace algunos años la utilización del diseño experimental se extendió al sector químico y agroalimentario, con un significativo un aumento de interés debido al desarrollo de la informática, que brindó herramientas y aplicaciones que facilitan su uso.

Glosario

Los términos que usualmente se emplean con mayor frecuencia en el diseño experimental son:

factor. Podría ser equivalente a “variable”. Se trata de cada uno de los aspectos que puede incidir en los resultados de un proceso o de un ensayo. Hay ***factores controlados*** que son aquellos sobre los que el operador tiene la posibilidad de modificar sus magnitudes y grado de injerencia en el proceso que examina. Existen ***factores sin control***, que son aquellos inmodificables o inaccesibles por parte del operador. Para comprender por qué no se usa el término variable vamos a usar un ejemplo: El tamaño o capacidad de una paila es una ***variable*** ya que se construyen y se pueden elegir en una gama muy amplia de capacidades. Sin embargo, una vez instalada, su tamaño es fijo. Si se necesita elegir una nueva paila, entonces el tamaño es un ***factor*** a tener en cuenta. Si, por el contrario, se trata del mejor uso de una paila existente, el tamaño ya es fijo.



A su vez, los ***factores sin control*** pueden ser

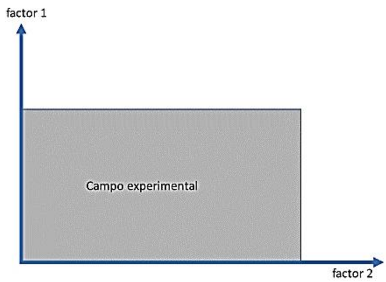
factores fijos (este aparente oxímoron se explica de la misma manera que en el caso anterior: la capacidad de una paila para elaborar dulce de leche es un ***factor variable*** ya que se pueden instalar de diferentes capacidades. Una vez instalada, en un ensayo de línea de producción, esa capacidad es un factor ***fijo***).

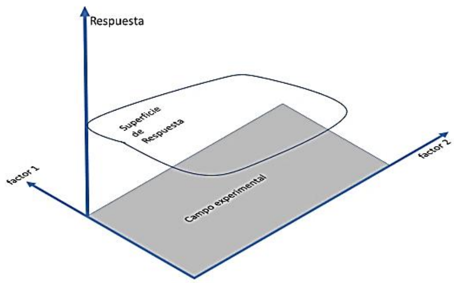
factores estocásticos, que son aquellas sobre las cuales el operador no puede ejercer control alguno en cuanto a la magnitud u oportunidad de sus cambios, ya sea porque son desconocidas o porque son muchas. En este grupo suele incluirse (o, directamente, ser) el error general del ensayo.

nivel. Se denomina así a cada uno de los estados de variación de un factor. Según el tipo de factor, los niveles pueden ser ***continuos*** o ***discretos***. En general se simplifican los niveles continuos llevándolos a puntos discretos simples. Por ejemplo, la temperatura de un proceso es un factor variable continuo (que quizá pueda variar entre 190 y 230 °C) pero suele llevarse a discreta haciendo que varíe en pasos de, por ejemplo, 10° o de 50° por vez o, mucho más simple, eligiendo tres niveles (190°C / baja; 210°C / moderada y 230°C / alta)

campo experimental. Es fácil ver que, si estamos experimentando con dos factores en dos niveles y si representamos los dos factores como ejes cartesianos ortogonales, obtendremos para cada una de las combinaciones de ambos un punto que será parte de una superficie que es el lugar geométrico de esos puntos. Ese es el campo experimental. Este ejemplo es simple porque son dos dimensiones, pero debemos tener en cuenta que ese campo para ***n*** factores, será ***n***—dimensional.

superficie de respuesta*.** Si agregamos un tercer eje cartesiano perpendicular a los del ejemplo del punto anterior y le asignamos ordenadas que representen la magnitud de la respuesta buscada en el experimento, tendremos que el campo experimental se convierte en una superficie de respuesta que, en este ejemplo, es tridimensional. Para ***n factores, será ***n***—dimensional.





¿Qué es el diseño experimental?

No existe una definición rigurosa para el término. Pero podemos aproximarnos si decimos que se trata de una herramienta basada en la inferencia estadística que tiene aplicaciones en diversos campos, aun en el campo de la propia herramienta. Sin embargo, existen algunas características importantes que permiten describir el concepto "***diseño experimental***". Así, entonces, el diseño de experimentos es un instrumento de

- ***caracterización.*** De los pasos y resultados obtenidos en un plan de investigación determinado.
- ***modelización***. Permite la generación de diversos tipos de modelos, algunos simultáneamente: un modelo de trabajo, un modelo predictivo

de comportamiento, un modelo final de descripción del fenómeno estudiado, un modelo de comportamiento del modelo de ensayo, etc.

- ***optimización***. Permite realizar optimizaciones en varios planos, nuevamente en algunos de ellos simultáneamente: optimizar el número de ensayos realizados, optimizar el número de variables y parámetros a estudiar, optimizar los procesos estudiados o los tratamientos disponibles, optimizar el modelo descriptivo utilizado, optimizar el criterio empleado en determinado tratamiento o proceso...

En las actividades de investigación y desarrollo hay que planificar una estrategia, para lo cual se establecen dos caminos posibles, el empírico y el determinista.

El método empírico, en principio, se relaciona con un modelo "caja negra", al cual se llega mediante experimentaciones e identificación de coeficientes.

El método determinista parte de la utilización de un modelo de conocimiento, con el que se validan las experiencias.

En todos los casos dispondremos de:

- información, cualitativa y cuantitativa,
- un método de optimización
- un comando, que es el control de la tarea.

La manera tradicional para planificar experiencias, en general, usa metodologías tales como:

- Hacer variar un factor a la vez, manteniendo los demás factores constantes. Podemos afirmar — y demostrar — que este proceso no permite obtener óptimos globales ni puede tener en cuenta las interacciones que pudieran existir entre factores. Esta manera de operar es conocida también mediante la locución latina *ceteris paribus* (literalmente "en igualdad de condiciones" se usa con el significado "permaneciendo lo demás en igualdad de condiciones")[3].
- Dividir el área experimental, lo cual aumenta el número de ensayos y genera dudas acerca de la precisión.

---

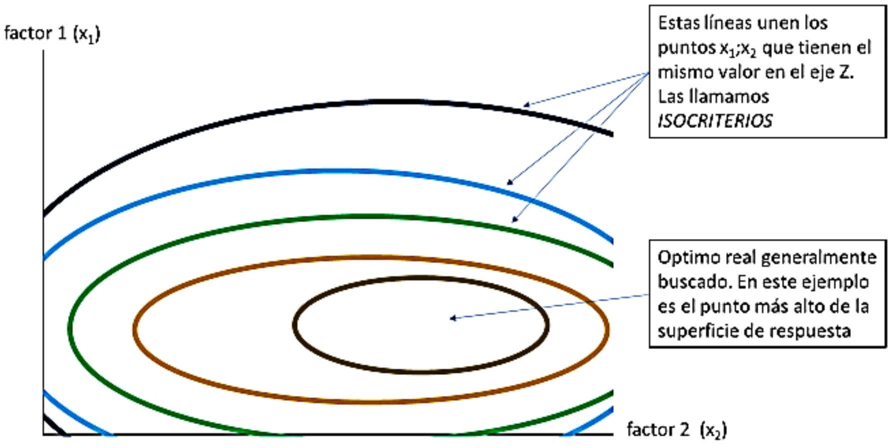
[3] Volvemos a analizar este método más adelante

En contraposición a estos métodos, los diseños experimentales estudian todos los factores simultáneamente y teniendo en cuenta las interacciones entre los factores, a la vez que incluye la forma de minimizar el número de experimentos necesarios.

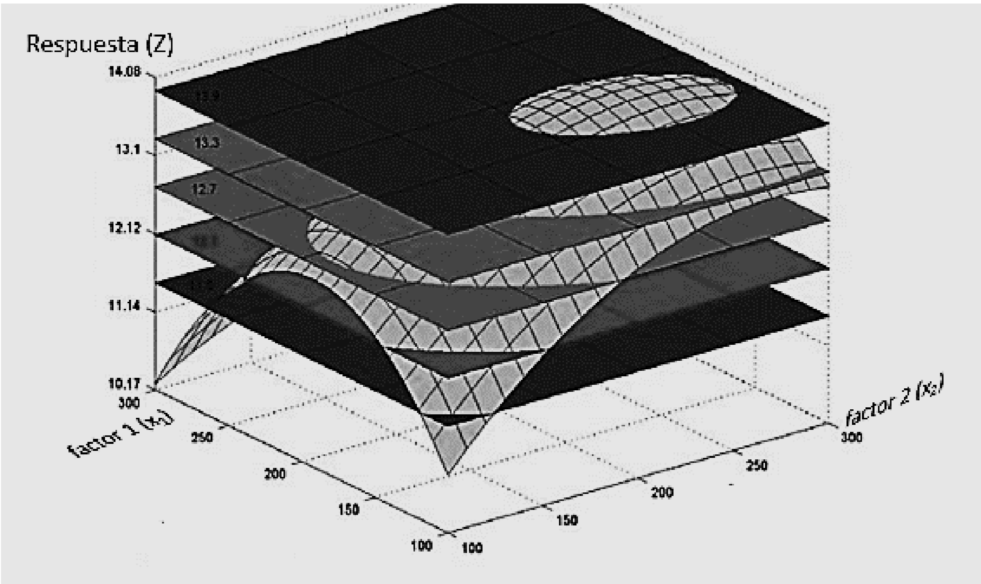
Métodos de búsqueda directa

Los modelos de búsqueda directa se emplean cuando no se conoce exactamente el comportamiento de los factores y la manera que ellos influyen en el objeto de estudio. Es un método empírico que emplea, de manera básica, un diseño de experimentos. Como ejemplo se muestra en forma esquemática un método de búsqueda directa llamado ***SIMPLEX EvOp***, que, si bien es un diseño atractivo, veremos — más adelante — conlleva ciertos riesgos.

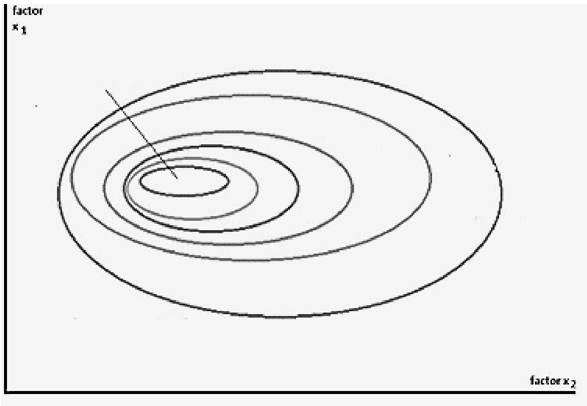
Si tenemos que hacer un ensayo variando dos factores a la vez, podremos suponer, como vimos más arriba, que ambos factores determinan una superficie de respuesta. Si graficamos los dos factores en un plano cartesiano ortogonal, cuyos ejes son cada uno de los factores, podremos representar el ensayo con una figura como la siguiente:



Con la figura de arriba presentamos algunos ***isocriterios*** en un hipotético caso de un efecto deseado (eje Z) que queda representado en un plano de dos dimensiones. Si ese mismo caso lo hubiésemos representado en tres dimensiones, lo veríamos como en la figura inferior:



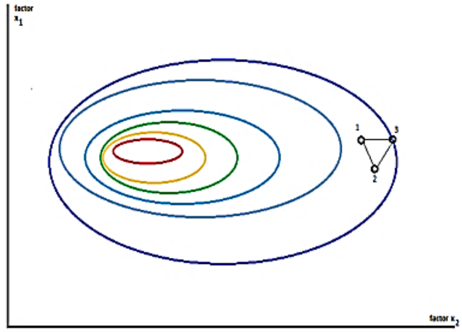
La primera figura representa un campo tridimensional visto en el plano de proyección de dos dimensiones determinadas por sendas variables x_1 y x_2. La tercera dimensión (la proyectada sobre el plano del dibujo) corresponde al valor que adquiere el criterio (Z) para cada una de las posibles — e infinitas — combinaciones [x_1 ; x_2], si $Z = f(x_1, x_2)$.

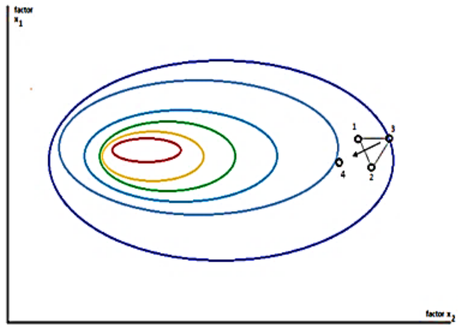


Para ver como se desarrolla un diseño con ***Simplex EvOp***, supongamos que tenemos el diagrama de isocriterios de la figura.

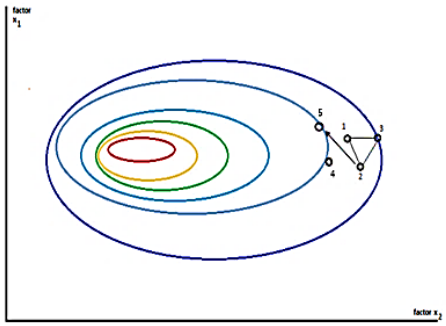
Obviamente, el investigador desconoce esta superficie de respuesta y debe encontrar, en este caso, el máximo en el eje Z haciendo ensayos con distintas combinaciones de x_1 y de x_2 (supongamos que x_1 es pH, que x_2 es concentración enzimática y que la respuesta es rendimiento). Este investigador debería hacer una enorme cantidad de ensayos para obtener algo parecido a la figura (que, recordemos, no conoce).

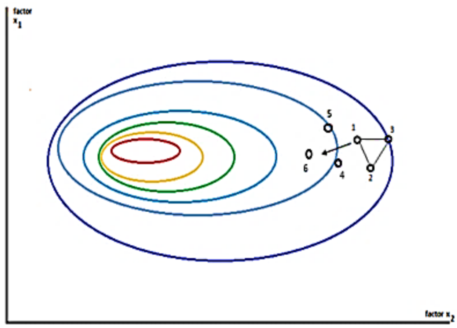
Usando el método Simplex EvOp va a realizar ensayos o experimentos en los puntos (combinaciones de pH y concentración.) numerados en las figuras que siguen. Comienza con tres ensayos, que corresponden a las combinaciones de las variables señaladas con los puntos 1, 2 y 3 elegidos de manera tal que sean los vértices de un hipotético triángulo. Al hacer los ensayos obtiene para cada vértice un valor del rendimiento (respuesta). El siguiente paso será reflejar ese triángulo según el vértice de peor respuesta (en el caso de la figura siguiente, el 3, que al reflejar define al 4). Así obtiene un nuevo triángulo y solo debe ensayar en la combinación definida por el vértice 4.

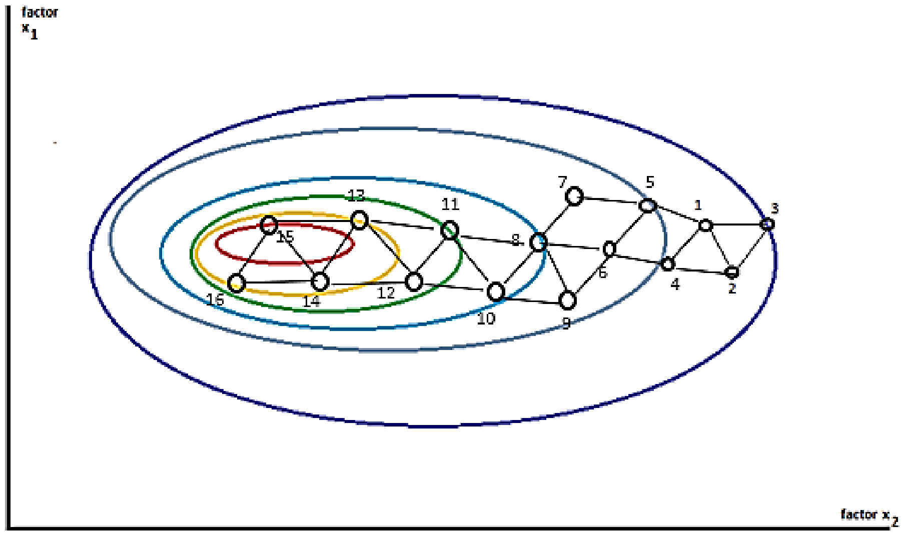




Del nuevo triángulo 1, 2 y 4, refleja el de peor resultado, el 2 que definirá el nuevo punto, 5) y así hasta que habrá un ensayo (el 15 en el ejemplo) que define un “punto calesita” alrededor del cual comienzan a repetirse ensayos. Ese punto se puede considerar como “***óptimo experimental***” porque no se está seguro de haber encontrado el óptimo real (ver figura, como ejemplo). Como se ve el método permite examinar cambios de ambas variables simultáneamente.



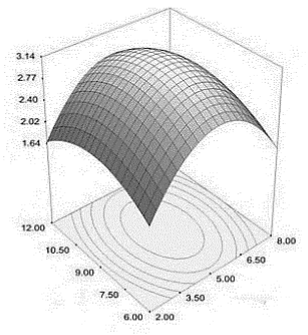


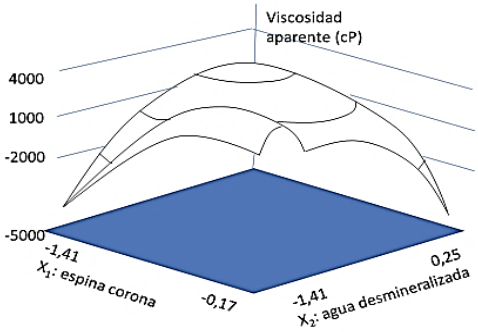


Los riesgos de SIMPLEX EvOp

Uno de los riesgos de utilizar el método Simplex EvOp es el de llegar a un óptimo local. Otro es la posibilidad de que nos encontremos con situaciones confusas si, por ejemplo, el paso que hemos elegido para hacer cada ensayo (paso es la diferencia de valor en cada factor X_1 y X_2 en la figura) es muy grande y “nos salteamos” el óptimo o, si, por el contrario, es muy pequeño y hacemos cantidades exageradas de ensayos.

En las figuras siguientes vemos ejemplos de ensayos utilizando superficies de respuesta



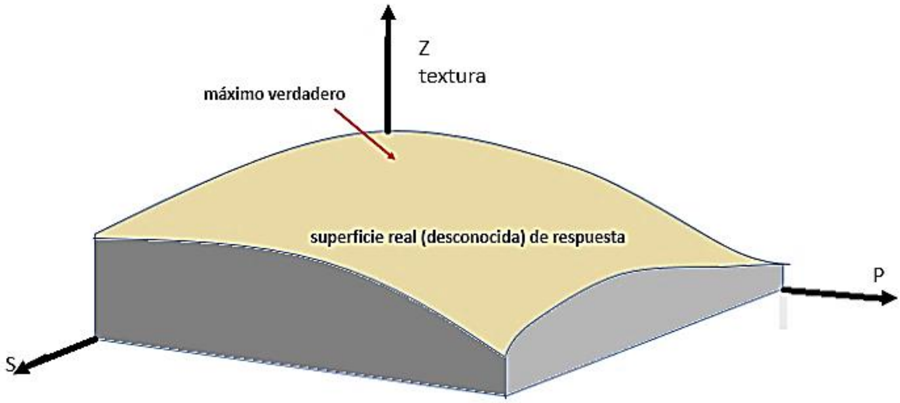


Los riesgos de variar un factor por vez

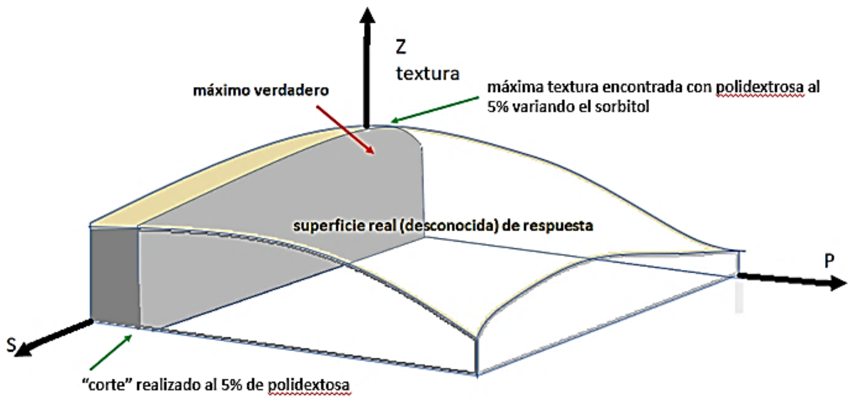
Vamos a ver ahora cuál es la diferencia entre este método y uno tradicional (en el cual el investigador mueve el valor de una variable por vez, *ceteris paribus*)

Es muy probable que, estudiando la variación de un factor por vez, nos encontremos con errores ligados al método. Por ejemplo, supongamos que queremos fabricar un bizcochuelo con harina y fibras dietéticas[4] y buscamos la mejor textura, sabiendo que depende de dos variables: el porcentaje de sorbitol (S) y el porcentaje de polidextrosa (P) que incorporemos a la mezcla.

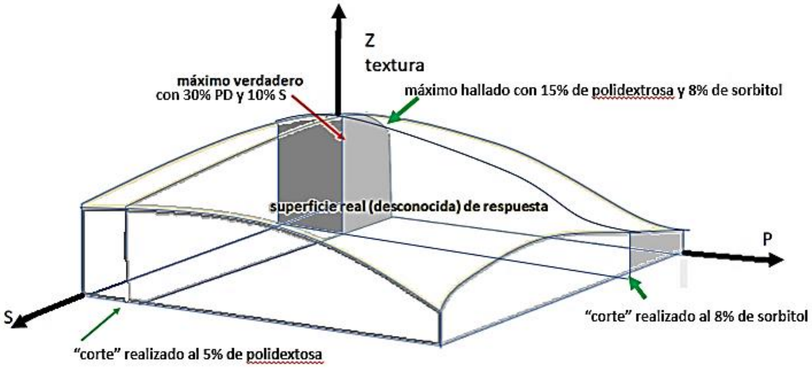
[4] Caracterización química y sensorial de bizcochuelos enriquecidos con fibra dietética y micronutrientes para el anciano. - Emma Wittig de Penna , Paula Avendaño, Delia Soto, Andrea Bunger. Facultad de Ciencias Químicas y Farmacéuticas. Universidad de Chile- Santiago. Chile



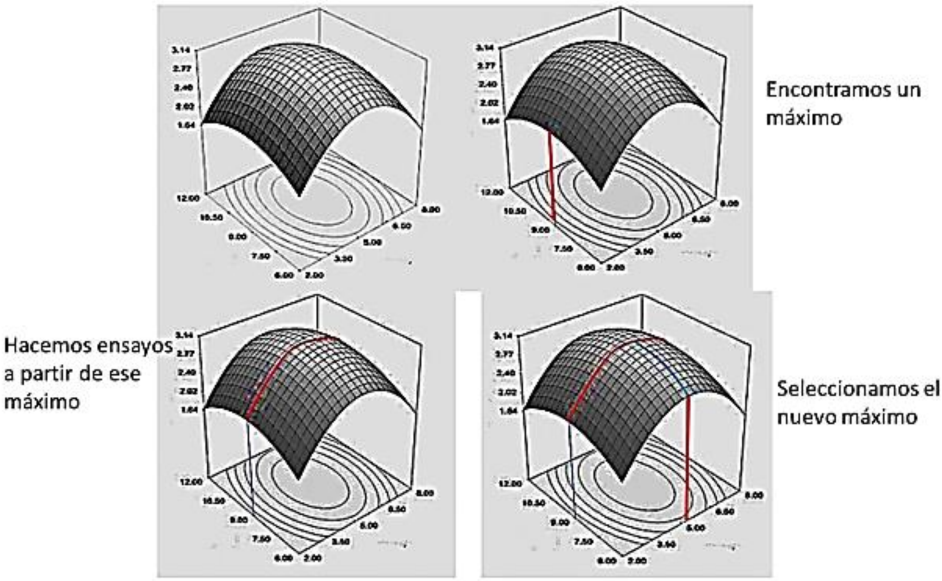
Si empleamos una forma tradicional de búsqueda, como *ceteris paribus*, podríamos mantener fija la formulación de Polidextrosa (por ejemplo, en 5 %) y variar el sorbitol hasta encontrar el máximo en la respuesta de textura, (supongamos que lo hacemos y que con 5 % de PD corresponde un porcentaje del 8 % de sorbitol).



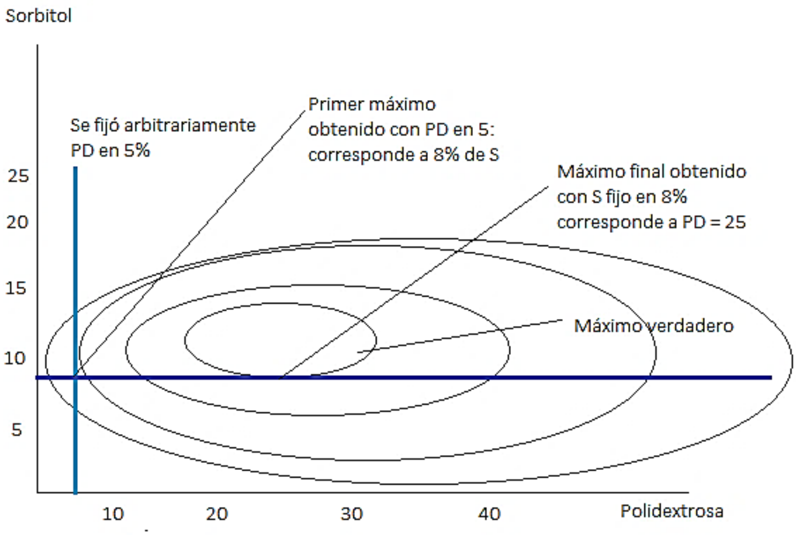
La variable que dejamos ahora fija es sorbitol al 5 %. Variamos ahora la variable polidextrosa hasta encontrar un máximo, la mejor textura (supongamos que lo encontramos al 15 %).



Informamos, entonces, este resultado (S = 8 %; PD = 15 %) como la combinación de factores óptima para encontrar la mejor textura. Aparentemente el método seguido es correcto y el resultado debería, por tanto, serlo. Pero, vemos en la figura de análisis, la elección de un punto inicial, en este caso al azar, (podría haber sido por ensayos preliminares o recetas en uso) nos lleva a un máximo relativo que no es el óptimo. En la figura de abajo vemos otro ejemplo de error cometido al varias un factor por vez:



En realidad, cuando la experiencia se hace variando ambos factores, mediante



un método de búsqueda directa adecuado, como el SIMPLEX EvOp, es más seguro encontrar el verdadero óptimo, sobre todo en casos como el de la figura que se encuentra en otra zona cercana, con coordenadas que corresponden a PD = 30 % y S = 10 %. Para esos valores, el método de fijar una variable y luego la otra tiene pocas probabilidades de coincidencia con la realidad, (depende del valor inicial arbitrario fijado para una de las variables a "congelar"), mientras que con SIMPLEX EvOp las probabilidades son más altas.

Diseños de experimentos variando todos los factores a la vez

Ahora bien ¿qué ocurre cuando se quieren estudiar varios factores? Como ya se mencionó, el mejor método es variar todos los factores a la vez. Si hay dos factores que determinan un resultado buscado, es evidente que corresponde un valor por cada uno de los infinitos valores de cada factor combinado con cada uno de los infinitos valores del otro.

En el caso de la textura, visto más arriba, si ésta se mide con una cantidad T que es función del % PD y del % S, (T = *f*(PD; S)) entonces habrá una textura ***Tn*** para cada par ordenado [PD_i ; S_j] ¿Quiere decir que hay que hacer un número ilimitado de ensayos que correspondan a las combinaciones posibles de cada valor de cada factor con las del otro?

Se podría pensar que, en la práctica, no es posible ensayar infinitos valores de PD y de S si estas son variables continuas. En primer lugar, **siempre hay extremos**, lo cual limita el campo de variabilidad: se sabe que el porcentaje de polidextrosa varía entre extremos más razonables que 0 y 100. Si por ejemplo se ensayara con pH, se sabe que hay valores que no son admisibles o que ciertos ingredientes no lo soportan, por lo cual se descartan esos extremos. Lo mismo puede afirmarse con la temperatura.

Una vez dentro de un rango, es usual elegir un valor para el "cambio", o un "paso" de la variable: se cambiará el porcentaje de polidextrosa o de sorbitol cada décima, cada unidad o cada centésima. Si trabajáramos con temperatura se variará de a un grado, de a medio o de a dos. Nuevamente el aparente número ilimitado ya no lo es, sino que sería razonablemente manejable, aunque aún puede ser muy grande.

Si adoptamos solamente estas previsiones podemos usar un diseño como SIMPLEX EvOp, efectivamente, partimos de un campo de variabilidad (extremos entre los cuales varían los factores) y de la elección de un "paso" o "salto" entre dos valores sucesivos de cada factor (distancia entre los vértices de cada triángulo).

Si no usamos SIMPLEX EvOp, en lugar de fijar un "paso" de variabilidad, podremos determinar un número fijo para la variación de los factores, a ese número de valores posibles de cada factor lo llamaremos **NIVEL**. El número de niveles puede ser muy pequeño (dos) o tan grande como con el sistema de pasos (diez, por ejemplo, si se decidiera explorar un rango de 10 con un paso de a una unidad por vez).

Supongamos, entonces, que decidimos que los dos factores variarán en dos niveles cada uno.

Por ejemplo, una determinada característica (un color deseable, por ejemplo) depende del **pH** y de la **temperatura** de horneado, en el mismo caso de los bizcochos. Ahora estableceremos los rangos de variabilidad de cada factor:

el **pH** puede ensayarse en 4 y en 6

la **temperatura** de horneado en 120 y en 150 °C:

Factor	Nivel bajo	Nivel alto
pH	4	6
Temperatura	120	150

Esto significa que programaremos los siguientes experimentos:

Experimento Nº	pH	Temperatura
1	4	120
2	4	150
3	6	120
4	6	150

Para aumentar la exactitud podríamos aumentar el número de niveles. También es probable que nos convenga trabajar con más factores que dos, ya que en la realidad los factores condicionantes son varios.

En cualquier caso, el número de experiencias que deberemos realizar está dado por:

$$N°de\ niveles^{N°\ de\ factores}$$

Por lo tanto, si queremos hacer un experimento con siete factores cada uno de los cuales lo analizaremos en cinco niveles, deberemos hacer $5^7 = 78125$ ensayos.

Generalmente es imposible planificar una secuencia de ensayos de este tamaño, debemos buscar, entonces, una alternativa que nos permita realizar menos experimentos, y llegar a un resultado igualmente confiable.

Las opciones para disminuir el número de experiencias pueden ser:

a. Disminuir el número de niveles, por ejemplo, pasar de

$$5^7 = 78125$$

$$3^7 = 2187$$

$$2^7 = 128$$

b. Disminuir el número de factores, por ejemplo:

$$5^7 = 78125$$

$$5^4 = 625$$

c. Combinar ambas alternativas, por ejemplo:

$$3^4 = 81$$

Existen procedimientos que nos permiten realizar estas búsquedas con la menor cantidad posible de experiencias. Para el caso de 7 factores a 5 niveles se realizan sólo 162 experiencias, con una programación que nos permita variar los niveles de todos los factores a la vez.

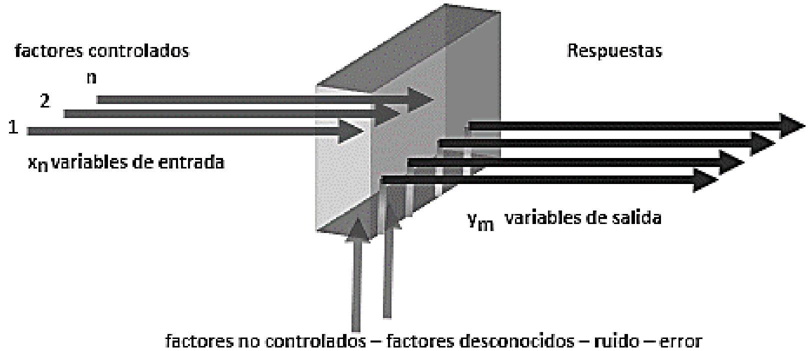
El objetivo común a estos métodos es:

- disminuir el número de ensayos
- detectar las interacciones
- detectar el o los óptimos
- modelizar los resultados

Métodos matemáticos y estadísticos

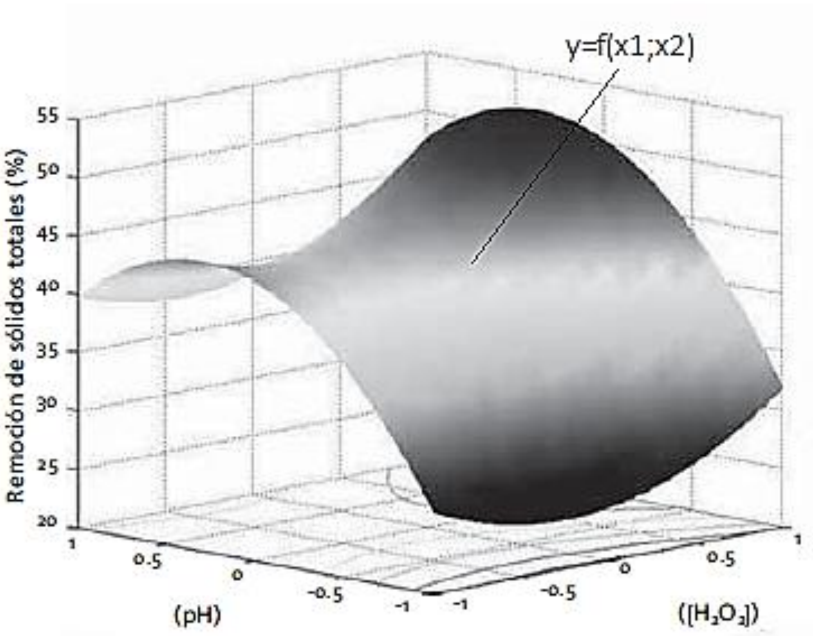
Utilizando una experimentación metódica lograremos realizar un mínimo de ensayos y obtener la máxima información posible. Estos métodos los usaremos para:

- Seleccionar los factores a estudiar. Para ello necesitamos evaluar los factores influyentes. Implica optimizar el método del diseño de experimentos que queremos optimizar.
- Estudiar un proceso. Transformación, formulación de productos. También para interpretar, dimensionar, regular y modelizar ese proceso.
- Optimizar un proceso.



El modelo global de comportamiento es:

$$Y = f(X_1; X_2; \ldots; X_n) + e$$



Donde ***Y*** representa el comportamiento de todos los factores controlados o que pueden ser controlados y que operan sobre el sistema como resultado del procesamiento de las variables de entrada ($\boldsymbol{X_j}$).

Aquellos factores imposibles de controlar o, aun, aquellos factores que no se conocen, se los incluye en el sistema. Así, nos encontraremos con una serie de variables que influyen sobre el sistema pero que no tenemos posibilidad de controlar. A esas variables se las conoce como “ruido.

De esta manera, el sistema opera mediante una función de cambio para cada factor que da como resultado un conjunto de variables de salida, que se denominan ***respuesta***. Este conjunto se obtiene operando cada variable de entrada por la función más el ruido, que, a los efectos de cálculo, puede ser pensado como ***error***. Otros términos que se utilizan son: ***superficie de respuesta*** (espacio que representa el conjunto de valores de la función $f(x_i)$, e “***isocriterios***” o “***isorespuestas***”).

Metodología del diseño experimental

Podríamos sintetizar la metodología del diseño de experimentos de la siguiente manera:

El primer paso, que aparenta ser trivial, no solo es imprescindible, sino que debe planificarse cuidadosamente. Solamente conociendo la mayor cantidad posible de factores y seleccionando de entre ellos cuáles son los que vamos a tener en cuenta, se puede elegir un método adecuado. Trabajar con muchos factores lleva a diseños pesados y confusos. Elegir pocos lleva el riesgo de dejar sin considerar un factor importante. Lo mismo será con la determinación de las respuestas a medir. **Entenderemos por respuesta al indicador que usemos para medir la influencia de los factores.**

También es importante establecer con toda claridad los límites del área experimental. Límites extensos pueden resultar en aumentos enormes del número de ensayos necesarios, muchos de ellos probablemente realizados en zonas donde se sabe que no hay posibilidades reales (por ejemplo, un pH extremo) o que dan combinaciones ya probadamente peligrosas (pH 7 y t=30ºC, por ejemplo).

Recién cuando efectivamente se conoce el campo experimental se podrá elegir el modelo de trabajo y realizar el diseño experimental.

Terminadas estas etapas, se realizarán los ensayos, siguiendo la programación que hemos hecho. De ello obtendremos datos, con los cuales intentaremos el ajuste al modelo y la determinación de los parámetros óptimos. Para que el modelo sea considerado válido, debe ser predictivo dentro de sus límites y dentro de los márgenes del error, que, a esta altura, debe estar cuidadosamente ponderado.

Lo siguiente será verificar la predicción que hacemos del óptimo del modelo. Eso lo hacemos mediante ensayos en el punto que el modelo pronostica que encontraremos ese óptimo. La cantidad de ensayos deberá ajustarse en función del error ya estimado.

Terminada estas etapas estaremos en condiciones de tomar la decisión de aceptar o rechazar la propuesta que nos hace nuestro propio modelo.

Ahora bien, este es un esquema global de un diseño experimental. Sin embargo, tengamos en cuenta que el propio diseño podría ser optimizado.

Noción de diseños óptimos

Frente a un programa de experimentos es necesario establecer dos objetivos, los que, en principio, parecen ser contradictorios:

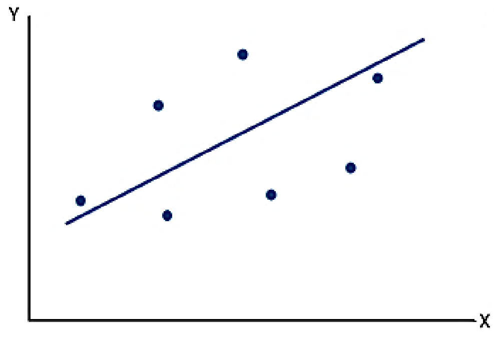
- el menor números de ensayos posible
- la máxima precisión en los resultados posible.

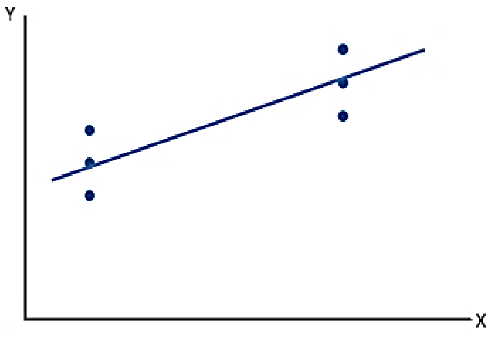
Siempre es deseable tener en cuenta a todos los factores en cada una de las experiencias, es por eso por lo que debemos trabajar en analizar todas y cada una de ellas cuando hacemos el diseño de experimentos, o, dicho de otra manera, programamos nuestros ensayos.

Localización de los puntos experimentales

Tomemos como base un diseño en dos niveles. Tendremos en cuenta que la recta (o cualquier función de respuesta que estemos buscando) podrá definirse mejor si los puntos que representan los datos están suficientemente alejados (separados) entre sí.

Por ello, los puntos experimentales deben estar en o cerca de los extremos del área de trabajo. La figura de la izquierda representa un ensayo que no respeta lo anterior, en contraparte a lo que muestra la de la derecha.





Diseño óptimo

Se llama diseño óptimo a aquel más económico que brinde la mayor información sobre los parámetros buscados:

- Mostrará la variación de TODOS los factores en cada experiencia
- Tendrá los puntos experimentales en los límites del área de trabajo.

Diseños factoriales completos a dos niveles.

Un diseño factorial COMPLETO es aquel que requiere realizar el número de experiencias, ***ex***, de acuerdo con:

$$ex = n^k.$$

Donde

n representa el número de niveles
k representa el número de factores

Este diseño presenta varias ventajas, entre ellas: es simple, requiere un número de experiencias relativamente bajo (esa cantidad dependerá de cómo se eligen los factores y los niveles), y, además, permite la elección de los factores.

Debemos tener presente que, si lo que buscamos son respuestas lineales, nos resultará imposible optimizar.

<u>Ejemplo</u>: Se estudia el rendimiento de una reacción química que depende de 2 factores: temperatura y presión, cada uno de ellos varía en dos niveles: bajo y alto. (Presión alta y baja, temperatura alta y baja)

Como es un problema de dos niveles y dos factores (2^2) realizaremos 4 ensayos o experiencias.

El primer paso será definir el valor de los factores y el de los niveles:

	Factor 1 (Temperatura) T	Factor 2 (Presión) P
Nivel – 1 (bajo)	60ºC	1 bar
Nivel +1 (alto)	80ºC	2 bar

Con estos datos vamos a construir la **MATRIZ DE EXPERIMENTOS:**

Número del ensayo	Factor 1 T	Factor 2 P
1	–1	–1
2	+1	–1
3	–1	+1
4	+1	+1

Respuesta η
60
70
80
95

La columna de la derecha indica el resultado que supuestamente hemos registrado al realizar cada uno de los ensayos.

La siguiente tabla que construiremos, y que llamaremos **MATRIZ DE LOS EFECTOS**, será:

Número del ensayo	Media	Factor 1 T	Factor 2 P	Interacción P—T
1	+1	–1	–1	+1
2	+1	+1	–1	–1
3	+1	–1	+1	–1
4	+1	+1	+1	+1
Divisor	4	4	4	4
Efectos	76,25	6,25	11,25	1,25

Respuesta
y_1=60
y_2=70
y_3=80
y_4=95

La columna marcada “*Media*” indica, para cada ensayo, el factor de ponderación respecto a la media de todos los niveles (en este caso, al ser uno, indica que todos los ensayos tienen el mismo “peso” o importancia relativa).

Vemos una nueva columna que representa cual es el grado de interacción entre los factores, en este caso, el 1 y el 2, cuyos valores se obtienen mediante el producto lógico entre los niveles correspondientes a cada ensayo: dos ensayos extremos producen una interacción máxima. Al contrario, dos ensayos antepuestos, por efectos compensatorios, producen la mínima interacción.

Con estos datos, para cada columna obtenemos un divisor, proveniente del número de ensayos evaluados, 4 en este caso para todos los factores.

Ahora calcularemos los diversos efectos:

- **Efecto de la media**

$$Em = \frac{1}{4}\left[(y_1 \times m_1) + (y_2 \times m_2) + (y_3 \times m_3) + (y_4 \times m_4)\right] = \frac{1}{4}\left[(60 \times 1) + (70 \times 1) + (80 \times 1) + (95$$

$$Em = 76{,}52$$

siendo m_i la media de cada ensayo *i*.

- **Efecto de la temperatura**

$$Et = \frac{1}{2}\left[\frac{1}{2}(y_2 - y_1) + \frac{1}{2}(y_4 - y_3)\right] = \frac{1}{4}(-y_1 + y_2 - y_3 + y_4) =$$

$$Et = \tfrac{1}{4}(-60 + 70 - 80 + 95) = 6{,}25$$

la diferencia $(y_2 - y_1)$ se utiliza para el cálculo del promedio, representa la diferencia de respuesta que se obtiene cuando la temperatura es baja y alta mientras la presión es baja.

De la misma manera, $(y_4 - y_3)$ representa las respuestas entre los dos niveles de temperatura cuando la presión es alta.

- **Efecto de la presión**

Usando los mismos criterios que en el punto anterior, tenemos:

$$Ep = \frac{1}{2}\left[\frac{1}{2}(y_3 - y_1) + \frac{1}{2}(y_4 - y_2)\right] = \frac{1}{4}(-y_1 - y_2 + y_3 + y_4) =$$

$$Et = \frac{1}{4}(-60 - 70 + 80 + 95) = 11{,}25$$

- **Efecto de la interacción Temperatura Presión**[5]

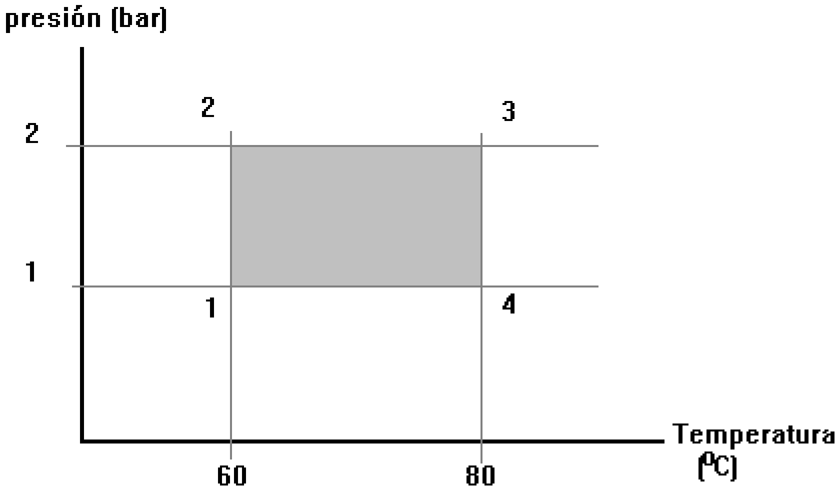
De la misma manera:

$$Etp = \frac{1}{2}\left[\frac{1}{2}(y_4 - y_3) + \frac{1}{2}(y_1 - y_2)\right] = \frac{1}{4}(y_1 - y_2 - y_3 + y_4) =$$

$$Etp = \frac{1}{4}(60 - 70 - 80 + 95) = 1{,}25$$

El modelo resultante, entonces, será:

Y = 6,25T + 11,25P + 1,25PT + 76,25



Desarrollo en hoja de cálculo

Podemos desarrollar este mismo ejemplo en una hoja de cálculo (***Excel*** o ***Calc***). Para eso, debemos recordar que en este caso no hay un método único, y podemos hacerlo de diferentes maneras.

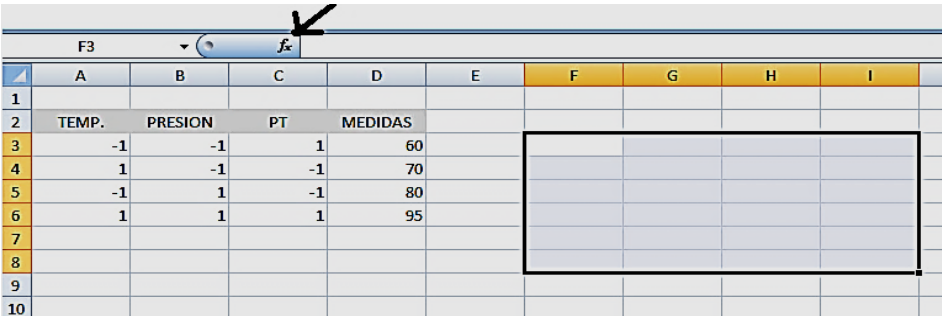
[5] Se analizó la interacción para presión alta y baja. Podríamos haber realizado el análisis para Temperatura alta y baja y aún podríamos hacer un promedio de ambas medidas (1,25; 0,75 y 1 respectivamente). En este caso las diferencias no son significativas frente al valor de los otros coeficientes hallados.

En este caso mostramos una de ellas, pero, como siempre, podemos ensayar otras. Primero vamos a plantear el problema escribiendo en una hoja, en forma ordenada los datos que tenemos.

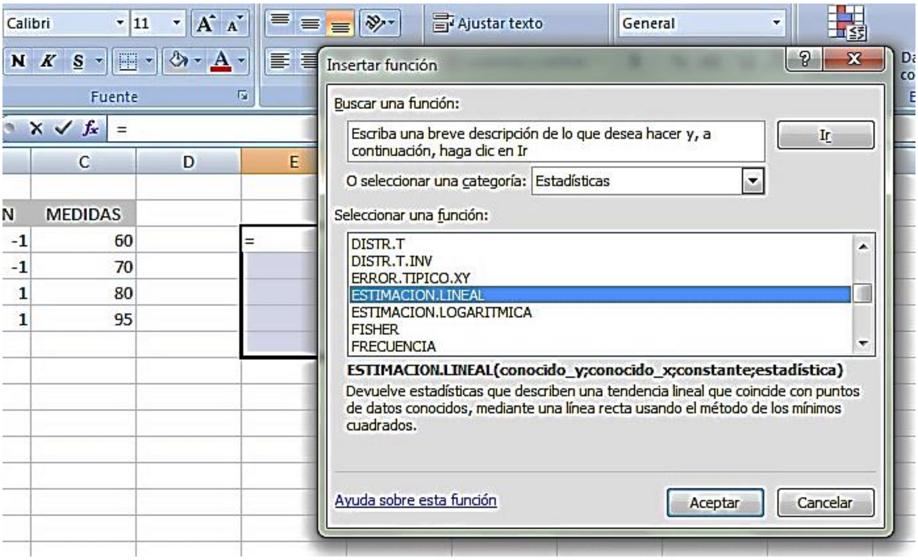
La figura siguiente nos muestra los datos cargados en un lugar que elegimos de esa hoja, que, repetimos, puede ser cualquier otro:

	A	B	C	D	E
1					
2	TEMP.	PRESION	PT	MEDIDAS	
3	-1	-1	1	60	
4	1	-1	-1	70	
5	-1	1	-1	80	
6	1	1	1	95	
7					

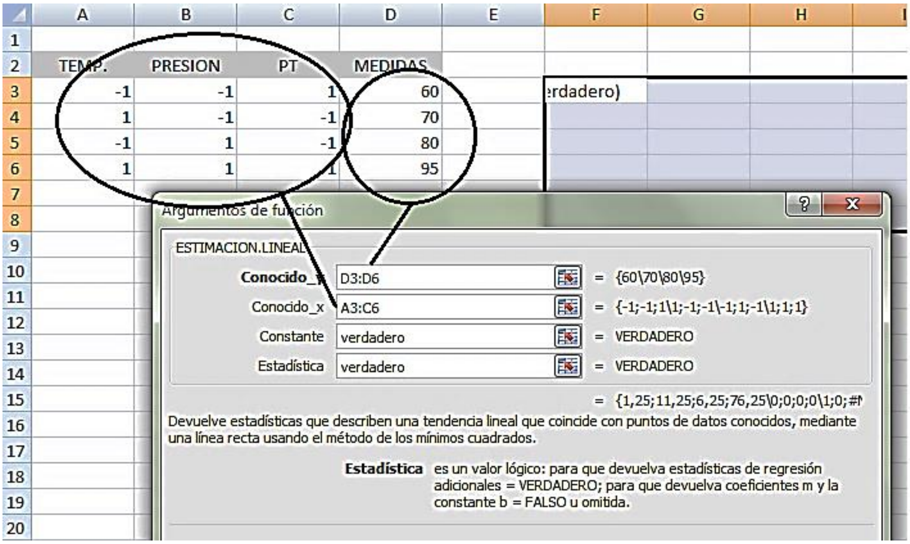
A continuación, buscamos un sector de la hoja que esté desocupado y allí seleccionamos un rango de cuatro columnas y cinco filas, de manera tal que ese conjunto de celdas en blanco quede "iluminado", y pulsamos el botón **[*fx*]**. Todo esto se muestra en la figura siguiente.



Va a aparecer un cuadro de asistencia de funciones. Donde dice "seleccionar la categoría" buscamos "*Estadísticas*" y, en listado de funciones estadística que veremos en la parte inferior, buscamos y seleccionamos la función "ESTIMACION.LINEAL"



Así estamos en condiciones de completar el cuadro de diálogo que aparece al aceptar, de la manera que ilustra la siguiente figura, asignando al campo "***Conocido_y***" el rango de resultado de los ensayos y al "***Conocido_x***" el rango de la matriz de los ensayos:



Ahora, como en toda función matricial, en lugar de pulsar "Aceptar" usaremos la combinación de teclas **CTRL-SHIFT-ENTER**, con lo que obtendremos, en el lugar que dejamos "iluminado", la siguiente matriz:

F3 | f_x {=ESTIMACION.LINEAL(D3:D6;A3:C6;VERDADERO;VERDADERO)}

	A	B	C	D	E	F	G	H	I	J
1										
2	TEMP.	PRESION	PT	MEDIDAS						
3	-1	-1	1	60		1,25	11,25	6,25	76,25	
4	1	-1	-1	70		0	0	0	0	
5	-1	1	-1	80		1	0	#N/A	#N/A	
6	1	1	1	95		#¡NUM!	0	#N/A	#N/A	
7						668,75	0	#N/A	#N/A	
8						#N/A	#N/A	#N/A	#N/A	
9										
10										
11										

Vamos ahora a interpretar la tabla obtenida:

1) si en el proceso de carga, en el campo llamado "*Constante*", escribimos, como lo hicimos, (VERDADERO) es que evaluamos la recta con la ordenada al origen ***b*** de la forma ***y = b + mx***. Si hubiésemos escrito (FALSO) o lo hubiésemos omitido (dejado en blanco), la recta evaluada sería ***y = mx***, recalculándose la (o las) pendiente(s) ***m***.

2) si en el campo "*Estadística*" ingresamos (VERDADERO), como lo hicimos, es que queremos calcular los estadísticos de regresión adicionales. Si lo omitimos o se ponemos (FALSO), solo se obtendrían los valores de las pendientes y, dependiendo de lo que hagamos con lo mencionado en el punto anterior, de ***b***.

3) la decodificación de la matriz resultante es la siguiente:

m_n	m_{n-1}	...	m_2	m_1	b
se_n	se_{n-1}	...	se_2	se_1	se_b
r_2	se_y				
F	d_f				
ss_{reg}	ss_{resid}				

I. ***m_j*** es la pendiente (o coeficiente de ***x_j***, para *j* = n hasta 1):
así, en el ejemplo:
m_3 = 1,25 (Interacción PT)
m_2 = 11,25 (Presión P)
m_1 = 6,25 (Temperatura T)

II. ***b*** es la ordenada al origen.
b = 76,25, en el ejemplo

III. ***se_j*** es el valor del error estándar correspondiente a cada ***m_j*.** El último valor será ***se_b*** que corresponde al error estándar de la ordenada al origen ***b*** (en el ejemplo es cero porque se usó la interacción. No quedaron márgenes para calcular ajuste)

IV. ***r_2*** es el coeficiente de determinación. Compara los valores calculados de ***y*** y los ***y*** teóricos. Varía desde 0 a 1. Si es 1, hay una correlación perfecta en la muestra, es decir, no hay diferencia entre el valor ***y*** calculado y el valor ***y*** teórico. En el otro extremo, si el coeficiente de determinación es nulo (0), la ecuación de regresión no es útil para predecir un valor de ***y***. (en el ejemplo tenemos que el ajuste es perfecto, por la interacción)

V. ***se_y*** es el error estándar de ***y***

VI. ***F*** es el valor observado del estadístico ***F***. Se utiliza para determinar si la relación observada entre las variables dependientes e independientes se produce por azar. En el ejemplo surge un error al dividir por cero (ya que no se dejaron grados de libertad para el error)

VII. ***d_f*** representa los grados de libertad. Se usan para encontrar valores ***F*** críticos en una tabla estadística y para determinar un nivel de confianza para el modelo

VIII. ***ss_{ref}*** y ***ss_{resid}*** muestra la suma de regresión de los cuadrados y la suma residual de los cuadrados

Ahora vamos a ver la pantalla que obtendríamos **sin usar la interacción** y de esa manera dejando grados de libertad para el error. Sería como la de la figura siguiente.

TEMP.	PRESION	MEDIDAS
–1	–1	60
1	–1	70
–1	1	80
1	1	95

11,25	6,25	76,25
1,25	1,25	1,25
0,99065421	2,5	#N/A
53	1	#N/A
662,5	6,25	#N/A

El error estándar es asumido por la interacción de las dos variables. Nos quedó un grado de libertad para medir el error y la bondad del ajuste es del 99%.

El modelo resultante es:

$$\boldsymbol{Y = 1{,}25\ PT + 11{,}25\ T + 6{,}25\ P + 76{,}25}$$

O bien, usando la segunda pantalla

$$\boldsymbol{Y = 11{,}25\ T + 6{,}25\ P + 76{,}25}$$

con un error estándar de 1,25

Uso del estadístico F y de los grados de libertad[6]

En el ejemplo que estamos siguiendo, vemos que — usando ESTIMACION LINEAL — el coeficiente de determinación, o $\boldsymbol{r_2}$, es 0,99065421 lo que indica que hay relación entre las variables y el rendimiento.

En este punto, podemos usar el estadístico ***F*** para determinar si los resultados, con este valor de $\boldsymbol{r_2}$, se podrían haber encontrado por otras causas no relacionadas con las investigadas o por casualidad.

---

[6] Este tema se completa con un desarrollo más detallado en el ítem “Prueba estadística del factor temperatura” más adelante en este capítulo.

Supongamos por un momento que, en realidad, no existe relación con las variables, pero que usamos una muestra que, por casualidad, hace que el análisis estadístico sea el que vimos. Vamos a usar el símbolo "$\boldsymbol{\alpha}$" para la probabilidad de llegar a la conclusión errónea de que hay una relación con las variables cuando en realidad no la hay y fue por casualidad.

Los valores ***F*** y ***df*** del resultado de la función ESTIMACION.LINEAL se pueden utilizar para determinar la probabilidad de que se produzca por azar un valor ***F*** más elevado.

Podemos comparar ***F*** con los valores críticos de las tablas de distribución ***F*** o podremos utilizar las funciones DISTR.F y DISTR.F.INV de Excel para calcular la probabilidad de que se produzca por azar un valor ***F*** mayor.

La distribución ***F*** apropiada tiene los grados de libertad ***v1*** y ***v2***. Si llamamos ***n*** al número de puntos de datos, entonces

$$\boldsymbol{v1} = \boldsymbol{n} - \boldsymbol{df} - 1$$

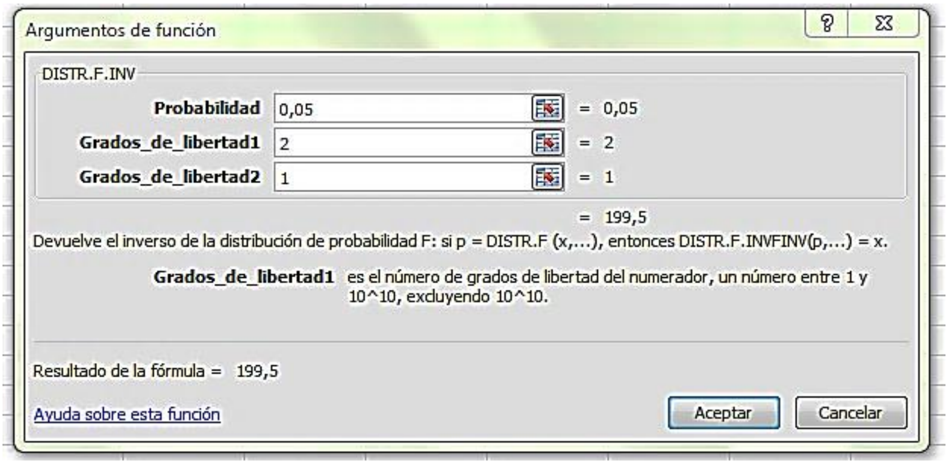
y

$$\boldsymbol{v2} = \boldsymbol{df}.$$

La función DISTR.F.INV (con la sintaxis DISTR.F.INV(p, v1, v2)) devolverá el valor crítico de F. Para eso, vamos a establecer los parámetros a usar en el cálculo de DISTR.F.INV.

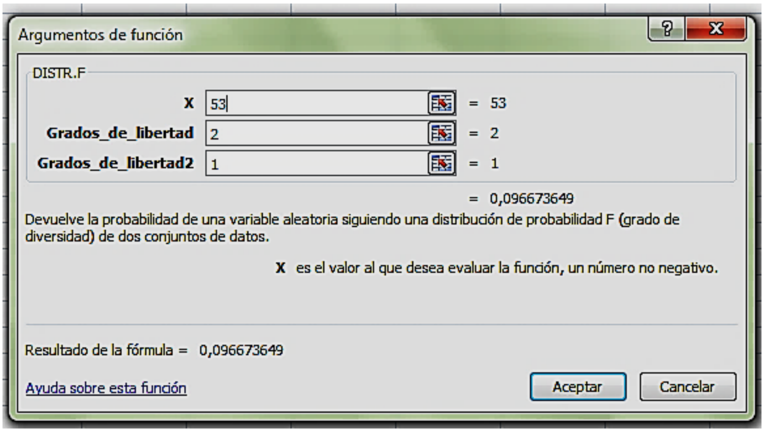
Podemos asignar un $\boldsymbol{\alpha} = 0{,}05$, o, dicho en otros términos, aceptamos una probabilidad de 0,05 de llegar a una conclusión errónea (95 % de seguridad). En ese caso, obtendremos, el siguiente conjunto de parámetros a ingresar en el cálculo de DISTR.F.INV:

$$\boldsymbol{\alpha} = 0{,}05$$
$$\boldsymbol{v1} = \boldsymbol{n} - \boldsymbol{df} - 1 = 4 - 1 - 1 = 2$$
$$\boldsymbol{v2} = \boldsymbol{df} = 1$$



obtenemos un valor crítico para ***F*** de 199,5. Como el ***F*** muestral (***F*** = 53) es mucho más bajo que 199,5, es probable que un valor ***F*** tan bajo se produzca por azar. (Con ***Alfa*** = 0,05, la hipótesis de que no hay relación entre ***Conocido_y*** y ***Conocido_x*** hay que rechazarla cuando ***F*** sobrepasa el nivel crítico, 199,5).

¿Cuál es la probabilidad de que se produzca por azar? Utilizamos DISTR.F para evaluar que probabilidad hay de que el valor de ***F*** (53) se haya obtenido por azar:



Vemos que la probabilidad encontrada es del 10%. Podemos llegar a la conclusión, ya sea buscando el nivel crítico de ***F*** en una tabla, o utilizando la función DISTR.F, sobre la ecuación de regresión, si es útil o no para predecir el valor del rendimiento de los procesos en función de la temperatura y la presión. Recordemos que es importante utilizar los valores correctos de ***v1*** y de ***v2***.

Diseños factoriales fraccionados o incompletos

Tomemos el siguiente ejemplo: Se desea obtener una emulsión con tres factores a dos niveles cada uno. La respuesta buscada es la estabilidad de la emulsión.

Si queremos hacer un ensayo factorial completo, necesitaremos realizar 8 ensayos, puesto que Nº de ensayos:

$$ne = 2^3 = 8.$$

Factor	Nivel Bajo (—)	Nivel Alto (+)
1-Ácido Graso	Baja Concentración	Alta Concentración
2-Ácido orgánico	Poca dilución	Mucha dilución
3-Tipo de emulsificante	A	B

Matriz de efectos del diseño completo:

Nº Ensayo	Media	Fact. 1	Fact. 2	Fact. 3	Inter. 1–2	Inter. 1–3	Inter. 2–3	Inter. 1–2–3	Resp
1	+1	–1	–1	–1	+1	+1	+1	–1	38
2	+1	+1	–1	–1	–1	–1	+1	+1	37
3	+1	–1	+1	–1	–1	+1	–1	+1	26
4	+1	+1	+1	–1	+1	–1	–1	–1	24
5	+1	–1	–1	+1	+1	–1	–1	+1	30
6	+1	+1	–1	+1	–1	+1	–1	–1	28
7	+1	–1	+1	+1	–1	–1	+1	–1	19
8	+1	+1	+1	+1	+1	+1	+1	+1	16
Divisor	8	8	8	8	8	8	8	8	
Efectos	27,25	–1	–6	–4	–0,25	–0,25	–0,25	0	

Si quisiéramos hacer menos ensayos podríamos acudir a un diseño incompleto o fraccionado. A continuación vemos la matriz de efectos para un diseño fraccionado obtenido a partir de la tabla anterior (en la sección siguiente se discutirá de qué manera se obtuvo el diseño reducido):

Nº Ensayo	I	Fact.1	Fact.2	Fact.3	Inter. 1–2	Inter. 1–3	Inter. 2–3	Inter. 1–2–3
5	+1	–1	–1	+1	+1	–1	–1	+1
2	+1	+1	–1	–1	–1	–1	+1	+1
3	+1	–1	+1	–1	–1	+1	–1	+1
8	+1	+1	+1	+1	+1	+1	+1	+1
Divisor	4	4	4	4				
Efectos	27,2	–0,75	–6,25	–4,25				

Hay que tener en cuenta que al realizar esta reducción encontramos un "costo oculto": si observamos detenidamente la tabla anterior vemos que en el conjunto de ensayos cada vez que el factor 1 pasa de (— 1) a (+ 1), la interacción 2—3 pasa también de (— 1) a (+ 1), por lo cual sería imposible saber si la variación en la respuesta obedece a la variación del factor 1 o a la variación de la interacción de los otros dos factores. Existen, por tanto, respuestas que pueden obedecer a efectos de factores o a efectos de las interacciones. Llamamos L_i a la respuesta y E_i al efecto y vemos los siguientes enmascaramientos:

L1=E1 + E23
L2=E2+ E13
L3 = E3 + E12
LM = I + E123

Cada uno de estos enmascaramientos se denomina ***alias***:

Para el factor 1:
su alias es la interacción del 2 y el 3

L1=E1 + E23

Para el factor 2
su alias es la interacción del 1 y el 3

L2=E2+ E13

Para el factor 3
su alias es la interacción del 1 y el 2

L3=E3+ E12

El modelo matemático para diseños $\mathbf{2^k}$ (2 niveles y k factores) es, para primer orden:

$$Y = b_O + \sum b_i X_i + \sum b_{ij} X_i X_j$$

donde b_o es la media,

b_i es el efecto del factor i

b_{ij} es el efecto de la interacción ij

Reducción o fraccionamiento

Existen varios métodos para obtener diseños fraccionados. En todo caso debe cuidarse de obtener un conjunto de experimentos que quede balanceado, o sea, que el número de cambios para todos y cada uno de los factores sea igual al de los demás. Esto restringe el campo de fraccionamientos, así en diseños 2^3 solo es posible un fraccionamiento (donde cada fraccionamiento queda con un número total de experiencias igual a la mitad del número de experiencias originales), en el de 2^7, se admiten cuatro fraccionamientos (donde cada uno tiene una octava parte del número original de experiencias).

Como ejemplo de fraccionamiento vamos a analizar el anterior, del cual solo tomamos la parte de la tabla que corresponde a los factores, sin las interacciones.

Reemplazando el nivel bajo (—) por cero (0) y el nivel alto (+) por uno (1), obtenemos dos grupos:

el primero será el grupo en que la suma de los renglones (o sea, la suma de los ensayos) arroja un número par

el segundo es el grupo remanente, o, lo que es lo mismo, aquellos renglones en que la suma arroja un número impar.

Vamos a sombrear los impares y dejamos sin sombra los pares:

Nº Ensayo	∑	Fact. 1	Fact. 2	Fact. 3					
1	0	0	0	0					
2	1	1	0	0					
3	1	0	1	0					
4	2	1	1	0					
5	1	0	0	1					
6	2	1	0	1					
7	2	0	1	1					
8	3	1	1	1					

Si seleccionamos el subgrupo de los “impares”, vemos que presenta cuatro ensayos, y en el conjunto cada factor está dos veces en el nivel bajo y las otras dos en el alto.

Nº de ensayo		Fact. 1	Fact. 2	Fact. 3
2	1	1	0	0
3	1	0	1	0
5	1	0	0	1
8	3	1	1	1

Si seleccionamos el de los “pares”, obtenemos un cuadro similar al anterior: cuatro ensayos, balanceados.

Nº de ensayo		Fact. 1	Fact. 2	Fact. 3
1	0	0	0	0
4	2	1	1	0
6	2	1	0	1
7	2	0	1	1

Cuando hay más factores se hacen “cortes” sucesivos, con criterios similares al anterior. Por ejemplo, separando los dos grupos de pares e impares, y sobre estos operar con algún otro criterio (división modular, por ejemplo).

También podremos realizar estas combinaciones usando hoja de cálculo, donde resulta muy cómodo escribir la tabla completa y, usando las funciones lógicas y

BUSCARV, encontrar la matriz reducida. El siguiente es un ejemplo para cuatro factores en dos niveles.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	**Cuatro factores a dos niveles: Diseño completo.**															
2		f1	f2	f3	f4	f1f2	f1f3	f1f4	f2f3	f2f4	f3f4	f1f2f3	f1f2f4	f1f3f4	f2f3f4	f1f2f3f4
3	1	-1	-1	-1	-1	1	1	1	1	1	1	-1	-1	-1	-1	1
4	2	-1	-1	-1	1	1	1	-1	1	-1	-1	-1	1	1	1	-1
5	3	-1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	1	-1
6	4	-1	-1	1	1	1	-1	-1	-1	-1	1	1	1	-1	-1	1
7	5	-1	1	-1	-1	-1	1	1	-1	-1	1	1	1	-1	1	-1
8	6	-1	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	-1	1
9	7	-1	1	1	-1	-1	-1	1	1	-1	-1	-1	1	1	-1	1
10	8	-1	1	1	1	-1	-1	-1	1	1	1	-1	-1	-1	1	-1
11	9	1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	-1	-1
12	10	1	-1	-1	1	-1	-1	1	1	-1	-1	1	-1	-1	1	1
13	11	1	-1	1	-1	-1	1	-1	-1	1	-1	-1	1	-1	1	1
14	12	1	-1	1	1	-1	1	1	-1	-1	1	-1	-1	1	-1	-1
15	13	1	1	-1	-1	1	-1	-1	-1	-1	1	-1	-1	1	1	1
16	14	1	1	-1	1	1	-1	1	-1	1	-1	-1	1	-1	-1	-1
17	15	1	1	1	-1	1	1	-1	1	-1	-1	1	-1	-1	-1	-1
18	16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
19																

La figura de arriba nos muestra una carga de la tabla completa con las interacciones calculadas como producto.

R	S	T	U	V	W
Diseño reducido			a:Método de reducción		
	f1	f2	f3	f4	Suma
1	0	0	0	0	0
2	0	0	0	1	1
3	0	0	1	0	1
4	0	0	1	1	2
5	0	1	0	0	1
6	0	1	0	1	2
7	0	1	1	0	2
8	0	1	1	1	3
9	1	0	0	0	1
10	1	0	0	1	2
11	1	0	1	0	2
12	1	0	1	1	3
13	1	1	0	0	2
14	1	1	0	1	3
15	1	1	1	0	3
16	1	1	1	1	4

En esta planilla se copió la parte de la tabla completa que corresponde solamente a los factores.

Con la función “Reemplazar” reemplazamos todos los “—1” por “0”.

Con SUMAR, en cada renglón, se obtuvo la columna suma. En ella seleccionamos aquellos renglones que son pares. (o sea, el número 1, el 4, el 6, el 7, el 10, el 11, el 13 y el 16)

Obtenemos un total de ocho renglones, de los 16 originales.

Diseño reducido. b:Matriz reducida

	f1	f2	f3	f4	f1f2	f1f3	f1f4	f2f3	f2f4	f3f4	f1f2f3	f1f2f4	f1f3f4	f2f3f4	f1f2f3f4
1	-1	-1	-1	-1	1	1	1	1	1	1	-1	-1	-1	-1	1
4	-1	-1	1	1	1	-1	-1	-1	-1	1	1	1	-1	-1	1
6	-1	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	-1	1
7	-1	1	1	-1	-1	-1	1	1	-1	-1	-1	1	1	-1	1
10	1	-1	-1	1	-1	-1	1	1	-1	-1	1	-1	-1	1	1
11	1	-1	1	-1	-1	1	-1	-1	1	-1	-1	1	-1	1	1
13	1	1	-1	-1	1	-1	-1	-1	-1	1	-1	-1	1	1	1
16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Utilizamos como valor de comparación el número de ensayos, y con la función BUSCAR en forma exacta (en FALSO), obtenemos la tabla incompleta. Se resuelve con ESTIMACION.LINEAL como fue visto.

Con este método aparecen celdas que tienen el mismo significado sobre la respuesta. Por ejemplo, si hay un cambio dado en la respuesta que sospechamos que fue originado por un cambio en el factor f1, no lo podemos asegurar, porque cada vez que cambia el factor f1 también cambia exactamente igual la interacción f2f3f4.

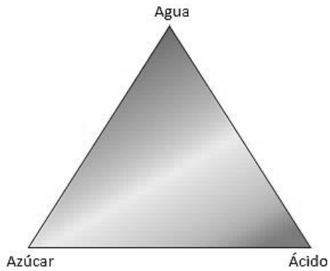
Los alias que se pueden encontrar en la planilla anterior son:

F1	F2F3F4
F2	F1F3F4
F3	F1F2F4
F4	F1F2F3

Diseños experimentales para más de dos niveles

Cuando hay que experimentar en más de dos niveles, los diseños factoriales son más difíciles de manejar en matrices codificadas como las vistas hasta aquí. Para hacerlo disponemos de varias alternativas, entre ellas:

- Diseño factorial 3^k
- Diseño rotable compuesto
- Diseño en red de Doehlert
- Área evolutiva
- Diseño de mezcla (responden a restricciones tipo $\sum X_i = 1$) y tienen la forma general que se ve en la figura siguiente.



La modelización matemática de estos diseños será

$$Y = b_o + \sum b_i X_i + \sum b_{ij} X_i X_j + \sum b_{ii} X_{i2}$$

donde el último monomio es el efecto de curvatura o efecto cuadrático.

Sin embargo, con estos métodos debemos tener en cuenta que es difícil visualizar tendencias asintóticas, fenómenos de saturación y óptimos locales.

Diseño rotable compuesto (Box-Wilson)

Es el método más clásico, se asocia a un modelo de segundo orden. El número de experiencias (mínimo) se encuentra con

$$ne = 2^k + 2k + 3$$

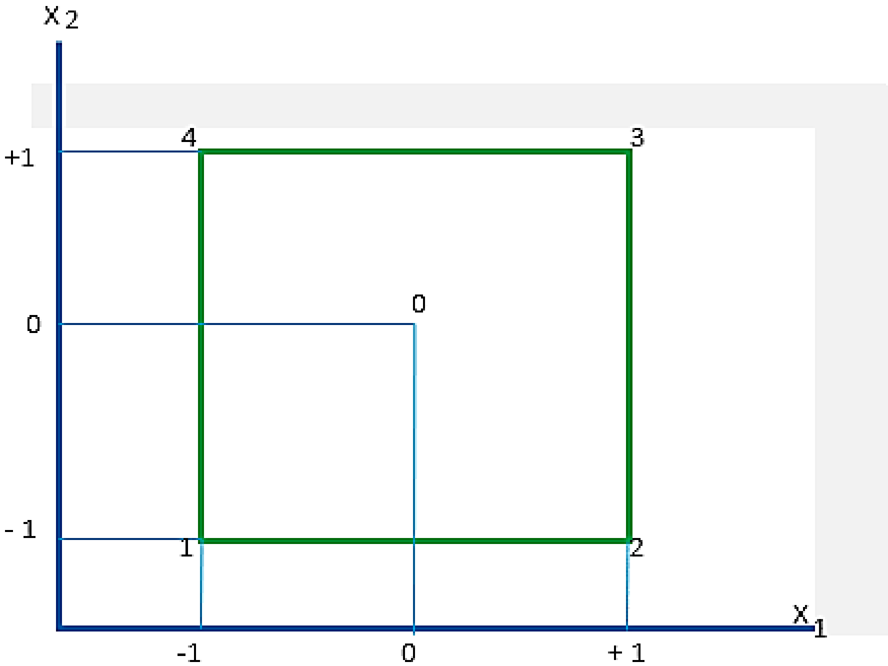
donde k es el número de factores.

Ejemplo en un área experimental de dos factores (área plana), en cinco niveles.

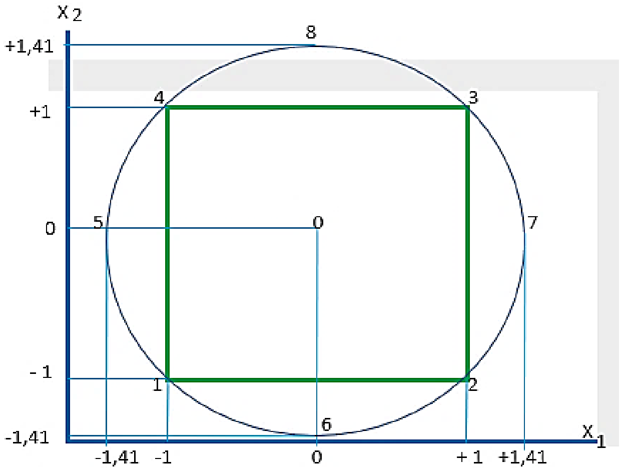
Los niveles originales son tres Un nivel bajo (–1), un nivel alto (+1) y un nivel intermedio o central (0), asociado al “punto conocido”. Este “punto central conocido puede ser la receta, la costumbre, lo especificado o el final de una búsqueda directa.

La figura de la página siguiente representa un diseño rotable compuesto para dos variables. Vemos la ubicación de los niveles “antes de rotar”, que es la operación que describiremos más adelante.

Estos primitivos tres niveles delimitan un área experimental “inicial” con forma de un cuadrado y cinco puntos de experimentación (el 1, el 2, el 3, el 4 y el central 0). Cada punto está definido por combinaciones nivel – factor.



En esos puntos se hacen los ensayos y así obtenemos datos experimentales que debemos analizar. Como este caso es un caso de búsqueda de optimización sobre superficie curva, un análisis estadístico de los resultados permite saber si se está en las proximidades del óptimo. De dar esto positivo, se “gira” el diseño y se encuentran nuevos puntos experimentales (puntos “estrella”) en niveles diferentes a los anteriores, así ahora hay otros cuatro nuevos puntos (el 5, el 6, el 7 y el 8 de la figura de la página siguiente) y haremos un nuevo ensayo repitiendo el punto central (0, que ahora es el ensayo 9). Esto permite entonces definir niveles diferentes a los del principio, que eran **– 1**, **0** y **+1**, agregando los dos niveles nuevos: el **–1,41** y el **+1,41**. (Raíz de dos, o sea del número de factores)



El mismo diseño de la figura anterior con los puntos "de rotación" marcados

En resumen:
El cuadrado representa los puntos del diseño factorial 2^2 niveles –1 y +1 (2^k)

1. (—1; —1)
2. (+1;—1)
3. (+1;+1)
4. (—;,+1)

El círculo representa los puntos en estrella "girados" o "rotados" del cuadrado 2 x 2 (2^k)

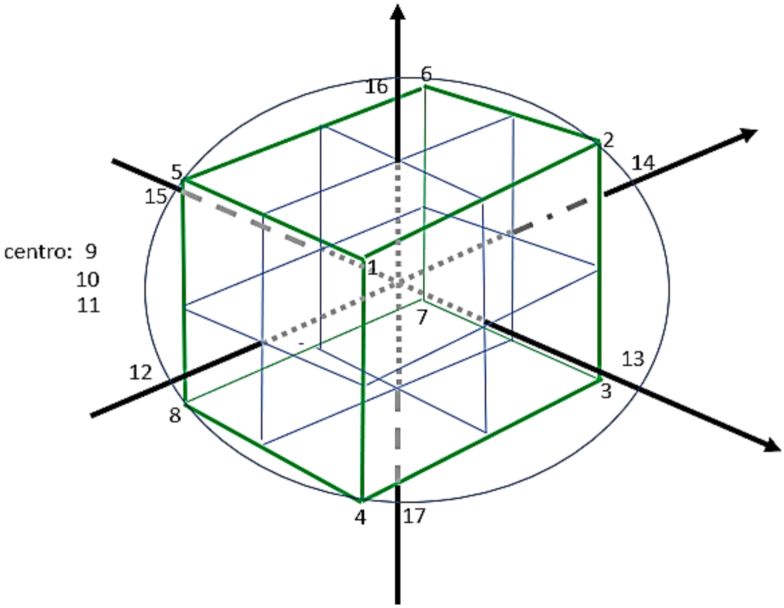
5. (—1,41 ; 0)
6. (0; —1,41)
7. (+1,41; 0)
8. (0; +1,41)

Por último, hay 3 repeticiones del punto central (sirven para testear el error experimental)

Así, el ensayo es 2^2 + 2x2 + 3 = 11 experiencias.

El mismo método, aplicado ahora a **<u>3 factores</u>**, nuevamente determina cinco niveles (tres niveles de partida y dos agregados por la rotación) daría estos resultados:

Niveles: —1,73; —1; 0; +1; +1,73 (1,73 = √3 o sea raíz del nº de factores)



Puntos factoriales:	2^3	(Nºs. 1 a 8)
Puntos estrella	2x3	(Nºs 12 a 17)
Puntos centrales	3	(Nºs 9, 10 y 11)
Total de experiencias	17.	

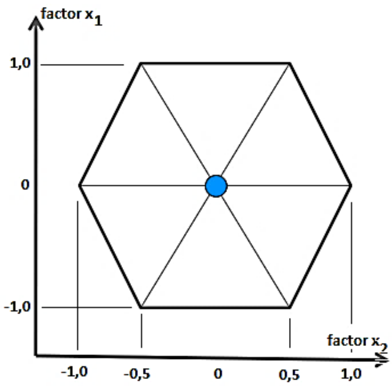
La que sigue es la matriz correspondiente a las experiencias en tres factores

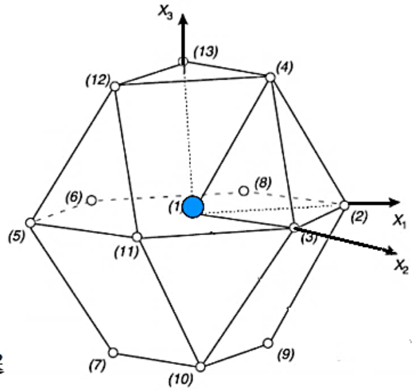
Prueba Nº	X1 (z)	X2 (y)	X3 (x)	
1	+1	–1	–1	
2	+1	+1	–1	
3	–1	+1	–1	DISEÑO
4	–1	–1	–1	
5	+1	–1	+1	
6	+1	+1	+1	FACTORIAL
7	–1	+1	+1	
8	–1	–1	+1	
9, 10, 11	0	0	0	PUNTO CENTRAL
12	0	–1,73	0	
13	0	0	+1,73	PUNTOS
14	0	+1,73	0	EN
15	0	0	–1,73	ESTRELLA
16	+1,73	0	0	
17	–1,73	0	0	

Si queremos obtener información más detallada sobre tratamiento estadístico y diseños Box Wilson, podemos consultar a partir de la página 60, más adelante

Matriz de Doherlet

También es un diseño asociado a un modelo de segundo orden, cuyo número total de experiencias es $\mathbf{k^2 + k + 3}$. Si tomamos como ejemplo una red a 2 factores: los puntos experimentales serán los tres clásicos: —1, 0 y +1 a los que luego se agregan — 0,866, — 0,5, +0,5 y +0,866.





Tanto el diseño rotable de Box Wilson como el de red tienen poder predictivo idéntico en todas las direcciones, junto a buena precisión pues la mayoría de los puntos están en los bordes experimentales y requieren pocas pruebas. La diferencia es que con Box Wilson se pueden estudiar en dos tiempos (antes y después de rotar) y con Doehlert se puede comenzar con pocos factores e incrementar luego.

Matriz de experiencias de una red de Doherlet con tres factores:

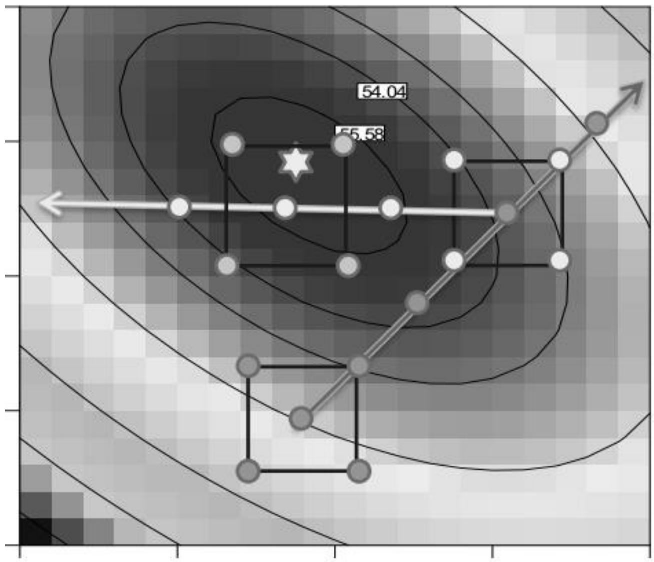
Experiencia Nº	X1	X2	X3
1	0	0	0
2	+1	0	0
3	–1	0	0
4	+0.5	+0,866	0
5	–0,5	–0,866	0
6	+0,5	–0,866	0
7	–0,5	+0,866	0
8	+0,5	+0,2887	+0,8165
9	–0,5	–0,2887	–0,8165
10	+0,5	–0,2887	–0,8165
11	0	+0,5774	–0,8165
12	–0,5	+0,2887	+0,8165
13	0	–0,5774	+0,8165
Nº de Niveles	5	7	3

Metodología del análisis de las superficies de respuesta

Recordemos que cuando mencionamos "superficie de respuesta" nos referimos a la representación gráfica de las variables como respuesta en el espacio de los factores. Aunque tenemos el inconveniente de lo difícil que resulta visualizar esas superficies cuando se trata de espacios de más de dos dimensiones encontramos la gran ventaja de que se identifican fácilmente las tendencias y es fácil localizar el óptimo. Puede requerir herramientas avanzadas de representación[7].

Hay varios métodos para recorrer experimentalmente la superficie de respuesta. Ya se mencionó el de ***Simplex EvOp*** y otras variantes, el de pendientes y tangentes, el de búsqueda por métodos como Fibonacci, etc.

La siguiente figura ejemplifica una búsqueda por el método de pendientes y tangentes en el que se arman diseños alrededor de aquellos puntos que no presentan diferencias significativas en las pendientes.



[7] Ver capítulo 21.

En este caso, el experimentador comienza haciendo los cinco ensayos determinados por el punto central y los cuatro vértices del cuadrado inferior central. El análisis de las pendientes indica que hay una tendencia a un máximo en el sentido de la flecha, lo cual se confirma con tres ensayos. El que da la respuesta máxima es el que se usa como nuevo punto central de una cuadrado de cuatro ensayos, con el que se repite el proceso. La nueva flecha indica el sentido de aumento, los tres nuevos ensayos y la elección un último cuadrado para confirmar que no hay pendientes significativas. Solo resta encontrar un método para determinar dónde está el máximo verdadero (señalado con una estrella en la figura).

Ejemplo de aplicación

Nos piden que analicemos un tratamiento por inmersión de productos animales. El objetivo consiste en estudiar la influencia de las variables del proceso sobre las transferencias de materia.

Factores fijos:

- agitación de la solución concentrada
- temperatura del baño maría
- dimensiones del producto

Factores a estudiar

• concentración de ClNa.	**Csal**	0 a 350 g
• logaritmo de la duración del tratamiento.	**ln(t)**	
• concentración de azúcar.	**Caz**	0 a 1900 g

Respuestas:

• Pérdida de agua	**Pa**
• Ganancia en ClNa	**Gsal**
• Ganancia en azúcar	**Gaz**

Elección del diseño:

Por ensayos preliminares sabemos que las respuestas no son lineales y que hay que considerar más de dos niveles por cada factor, por lo tanto, se elige una red de Doherlet con 15 ensayos:

$3^2 + 3 + 3$ repeticiones del punto central = 15 ensayos.

Con los datos haremos un análisis de regresión múltiple, cuya tabla de efectos de los factores y nivel de significación es la siguiente:

COEFICIENTES	Pa	Gsal	Gaz
lineales			
b_1	7,73 ***	1,66 **	–1,09 *
b_2	17,66 ***	–2,18 ***	2,41 **
b_3	16,68 ***	0,95 ***	4,09 ***
cuadrático			
b_{11}	–0,60 *	–0,46 .	0,76 .
b_22	–15,41 ***	1,79 **	–3,13 **
b_33	1,03 **		1,33 *
interacción			
b_12	–9,02 ***	–1,78 **	
b_{13}	3,26 **	0,83 *	
b_{23}	10,06 ***	–0,83 .	–2,42 *
constantes			
b_0	28,45	1,73	5,99
R^2	0,9957	0,9765 .	0,9601

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_11X_{12} + b_{22}X_{22} + b_{33}X_{32} + b_12X_1X_2 + b_{13}X_1X_3 + b_{23}X_2X_3$$

donde X_1 representa concentración de sal
X_2 representa concentración de azúcar
X_3 representa el ln del tiempo
*** nivel de significación p<0,1 %
** nivel de significación p< 1%
* nivel de significación p<5%

Espacios en blanco: coeficientes no significativos.

Conclusiones

Utilizar diseños experimentales y búsquedas directas es disponer de herramientas que constituyen un medio idóneo para caracterizar y optimizar procesos y formulaciones. No deberían encararse estas actividades sin tenerlas en cuenta y sin aprovechar las ventajas que brinda la posibilidad de utilizar en consecuencia las herramientas estadísticas. De todas maneras, hay que tener presente que tienen algunos límites:

- resulta difícil encontrar óptimos locales
- hay que limitarse al área experimental, las extrapolaciones son imprecisas.

Por otra parte, esas limitaciones se compensan frente a ventajas importantes, entre ellas:

- son técnicas versátiles que tienen la disponibilidad de diferentes diseños experimentales y de poderosas herramientas estadísticas
- se pueden optimizar tanto una respuesta simple como múltiples respuestas conjuntas
- requiere confirmación experimental de las predicciones y buen criterio del experimentador

CAPÍTULO 20. DISEÑO DE EXPERIMENTOS (2da. parte): Manejo estadístico de datos obtenidos en ensayos

Es inevitable que en algún momento tengamos a nuestra disposición una serie de datos experimentales que llegaron a nuestras manos quizá mediante planillas de operación, quizá sean los resultados de ensayos factoriales o de ensayos rotables o de búsquedas directas. En no pocas oportunidades, los datos pueden provenir de grandes encuestas o de recolecciones automáticas. Cualquiera sea el origen de esos datos experimentales, nuestra tarea será analizarlos y sacar conclusiones para la toma de decisiones. Para poder hacerlo, necesitaremos de herramientas estadísticas, entre ellas, análisis de la varianza y de regresión.

Frente a los datos, nuestra intención será encontrar un modelo que nos permita expresarlos singularmente. Veremos cómo. Comenzaremos a partir del modelo más simple, un modelo lineal.

$$z = b_0 + b_1 x$$

en este caso tenemos que determinar los valores de b_0 y de b_1, lo que podemos obtener con solamente dos niveles para la variable *x*. Estos ensayos deben repetirse para poder discriminar el error experimental. Como se ve (a diferencia de los que ocurría, por ejemplo, en el modelizado visto en programación lineal) en este caso debemos determinar los valores de los coeficientes de la función y se disponen los de las variables, mientras que en programación lineal se disponían los valores de los coeficientes y la tarea era determinar el valor de las variables.

El modelo general de arriba puede ser escrito de manera equivalente como

$$z = \mu + \alpha_i + e_{ij}$$

(Ecuación 1)

donde μ es el valor medio del ensayo,

α es la incidencia del factor utilizado

e es el error experimental

siendo i = 1, 2, ..., m (réplicas)

j = 1, 2, ..., n (niveles)

Si se llama

γ_{ij} al resultado de una corrida en el nivel j durante la repetición i

y llamando

γ** al resultado promedio de todos los ensayos *n.m*

y

γi* al resultado promedio de todos los *m* ensayos realizados en el nivel *j*

tendremos

$$\gamma ij = \gamma^{**} + (\gamma i^{*} - \gamma^{**}) + (\gamma ij - \gamma i^{*})$$

donde cada uno de los términos representa, respectivamente a μ, α_i y e_{ij}.(ver Ecuación 1)

De esta manera ahora vamos a analizar cómo podemos obtener un modelo de regresión sobre la base de una serie de datos. Para eso, vamos a ver el siguiente ejemplo:

Se realiza un ensayo con un solo factor, **temperatura**, a un nivel bajo (105ºC) y a nivel alto (110ºC), con tres repeticiones. Los resultados obtenidos (eficiencias) se muestran en la siguiente tabla:

		j = 1	j = 2
		Nivel 0	Nivel 1
i=1	Repetición 1	79	90
i=2	Repetición 2	80	91
i=3	Repetición 3	81	89

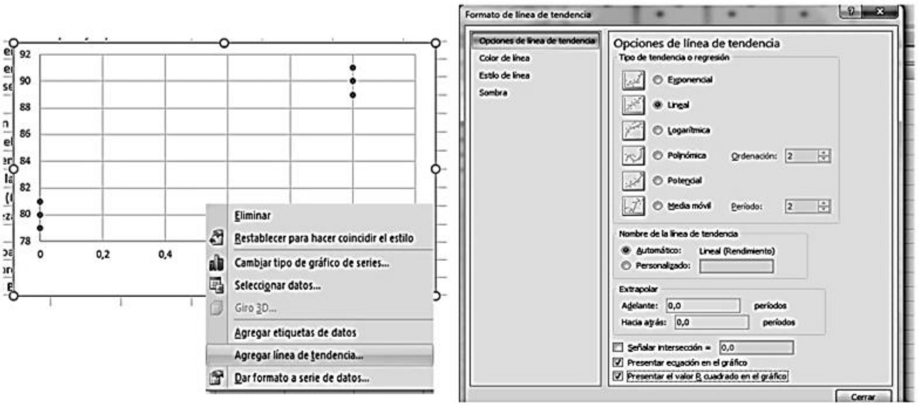
En primer lugar, reordenamos los datos a fin de calcular el modelo lineal y su coeficiente de regresión:

Nivel	Xij
0	79
0	80
0	81
1	90
1	91
1	89

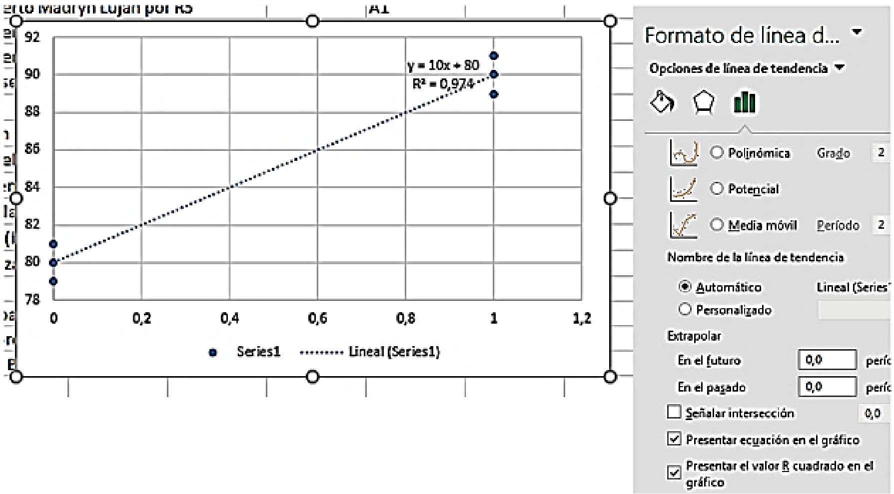
Mediante una planilla de cálculo podemos obtener el gráfico al que le adicionaremos el análisis de regresión correspondiente. (La curva de regresión calculada aparece en línea gruesa). El procedimiento es el siguiente:

1. cargamos los datos en una hoja de cálculo, en la forma mostrada en la tabla de arriba (Columna **Nivel** y columna factores **Xij**)
2. construimos el gráfico con la opción ***"XY – Dispersión"***, con el formato "puntos unidos por líneas".

3. Una vez terminado, editamos haciendo doble clic en el área de gráfico y hacemos un clic sobre la línea que une los puntos (seleccionamos la línea).
4. Con la línea seleccionada elegimos el menú “Agregar línea de tendencia...” (ver figura siguiente, abajo en el centro)
5. Ahora optamos por “Regresión Lineal” en la solapa “Opciones...” (ver figura siguiente, a la derecha)



6. Y por último marcamos las casillas “presentar ecuación en el gráfico” y “presentar el valor de R cuadrado...”



Con estos pasos obtenemos entonces el gráfico y su modelo lineal:

Y = 10X + 80
R2 = 0,974

que, llevado a la terminología empleada para este caso, significa:

$b_0 = 80$

$b_1 = 10$

Prueba estadística del factor temperatura

Intentaremos probar, mediante análisis de varianza, que la variabilidad del factor α_i es significativa frente a la variabilidad del error e_{ij}.

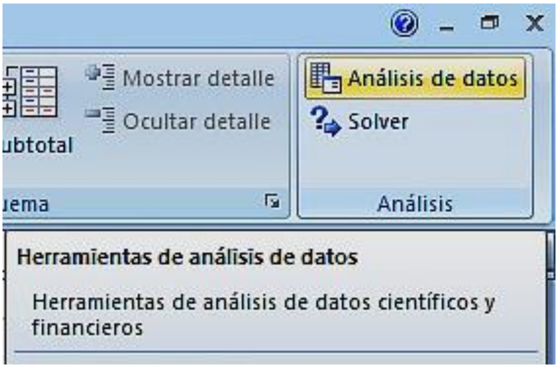
Además, mediante análisis de la regresión, debemos probar que la pendiente $b_1 = 10$ no admite el cero como solución ($\beta \neq 0$).

Análisis de varianza

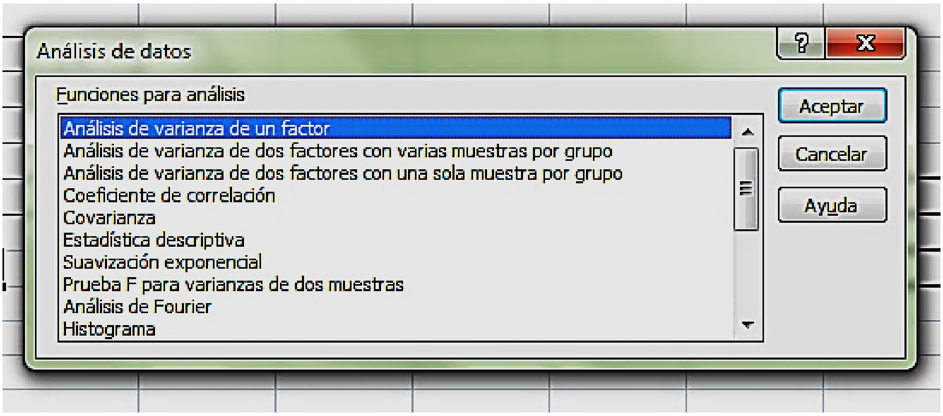
Para realizar el análisis de varianza con hoja de cálculo, vamos a partir de los datos que tenemos en la planilla, agrupados ahora por niveles de ensayo, como vemos en la figura que sigue:

	A	B	C	D	E	F
1						
2				j=1	j=2	
3				Nivel 0	Nivel 1	
4		i=1	Repetición 1	79	90	
5		i=2	Repetición 2	80	91	
6		i=3	Repetición 3	81	89	

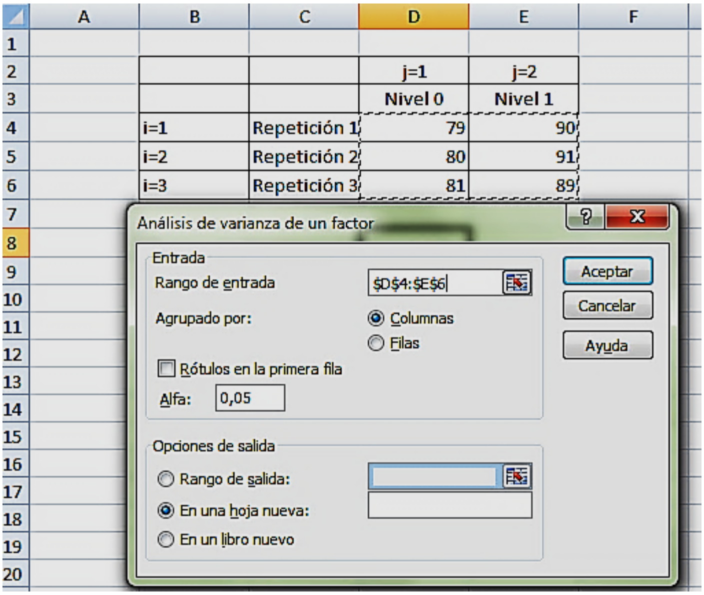
1. Seleccionamos en el menú Herramientas, el submenú "Análisis de datos..."



2. Y buscamos la opción Análisis de varianza para un solo factor.



3. Por último, seleccionamos el rango de datos (desde la celda que contiene “Nivel 0” a la celda que contiene el “89”; D3 a E6, en el caso de la figura)



Obtenemos una nueva hoja, en la que aparece ordenadamente toda la información del procedimiento de análisis de varianza que hemos hecho. El aspecto es similar al de la tabla que sigue:

RESUMEN

Grupos	Cuenta	Suma	Promedio	Varianza
Nivel 0	3	240	80	1
Nivel 1	3	270	90	1

ANÁLISIS DE VARIANZA

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Probabilidad	Valor crítico para F
Entre grupos	150	1	150	150	0,00025522	7,70864972
Dentro de los grupos	4	4	1			
Total	154	5				

En la fila marcada “*Entre grupos*” aparecerá lo que nosotros denominamos “variabilidad del factor” α, mientras que en el renglón “*Dentro de los grupos*” leeremos el error, la variabilidad del error.

La suma de los cuadrados será, entonces,

la expresión $(\gamma_{i*} - \gamma_{**})^2$ para la variabilidad del factor (*Entre los grupos*)
y la expresión $(\gamma_{ij} - \gamma_{i*})^2$ para el error (*Dentro de los grupos*)

Al ser 6 ensayos, disponemos de 5 grados de libertad. El factor tiene un grado y los restantes se asignan al error.

Los cuadrados medios surgen de dividir la suma de cuadrados por los grados de libertad.

El estimador ***F*** se calcula dividiendo los cuadrados medios del factor por los del error. Este valor es clásicamente obtenido en tablas de tres entradas: grados de libertad del numerador, grados de libertad del denominador y probabilidad.

La hipótesis nula (***H0***) será que las medias de ambos grupos son iguales, por lo tanto, la hipótesis a prueba es que no lo son, para ello, se debe cumplir que el valor esperado del numerador sea mayor que el del denominador.

Esto es, para ***H0*** falso, debe haber un valor grande de ***F***. Llamando a la probabilidad de error **p** y haciendo **p** = 1 — α.

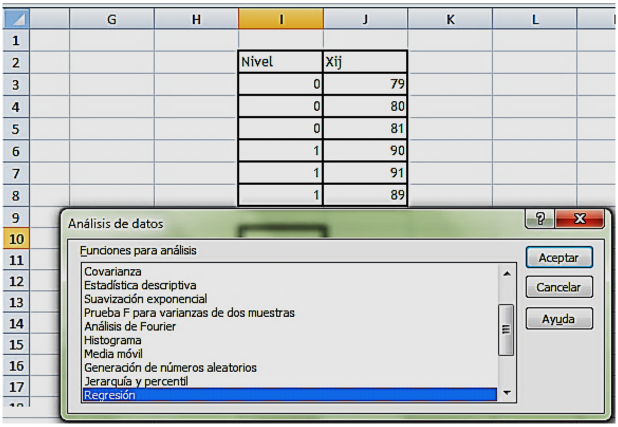
Se considera que la diferencia entre medias es

poco significativa si α = 0,05 o menos,
significativa si α = 0,01 o menos y
muy significativa si α = 0,001 o menos.

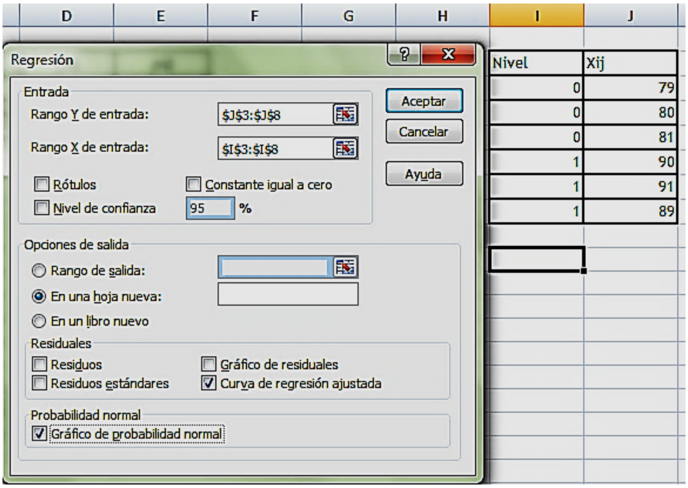
En este caso, la tabla de Análisis de Varianza presenta un valor de ***F*** de 150 (150/1), con un valor crítico para ***F*** de 7,7 y una probabilidad (de error) de 0,0002552, que es menor a α = 0,001, por lo que el valor de ***F*** resulta ***muy significativo***, o, en otras palabras, la probabilidad de la hipótesis nula es muy baja.

Análisis de regresión

Por otro lado, se puede realizar un análisis de regresión a fin de demostrar que la pendiente b_1 no admite el cero como solución, en forma similar al análisis de varianza. Para ello, se selecciona la opción "*Regresión*" en el cuadro de diálogo obtenido en el menú "*análisis de datos...*"



Los datos por analizar se obtienen de la tabla encolumnada "Nivel", "X_{ij}" que habíamos hecho primero. Ponemos como "Rango Y de entrada" la columna X_{ij} y como "Rango X..." la columna "Nivel":



Obtendremos los resultados que vemos en las tablas que siguen (varios de ellos los habíamos encontrado desde el comienzo del ejemplo). Pero de esta manera los disponemos en forma ordenada y ya tabulada.

La primera tabla indica los coeficientes de regresión, el número de observaciones, y el error típico. La segunda tabla es la que más interesa: vemos que es por completo similar a la de análisis de varianza ya vista, pero que difiere en las fuentes de variación.

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0,98692754
Coeficiente de determinación R^2	0,97402597
R^2 ajustado	0,96753247
Error típico	1
Observaciones	6

ANÁLISIS DE VARIANZA

	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	F	Valor crítico de F
Regresión	1	150	150	150	0,00025522
Residuos	4	4	1		
Total	5	154			

Comprobamos que, efectivamente, la primera fila ahora se denomina "*Regresión*" y representa a la Hipótesis en prueba ("*La pendiente es diferente de cero*"). Otra vez la probabilidad de error merece alta confianza (***H0*** es poco probable).

Complemento sobre ensayos factoriales

Vamos a completar lo que ya hemos visto sobre ensayos factoriales comenzando por resaltar que, para lograr diseñar un ensayo factorial debemos

1. elegir un número adecuado de corridas

2. fijar un número adecuado de factores (en el ejemplo que sigue, 3 factores)
3. elegir un número adecuado de niveles
4. determinar la separación o paso entre niveles

Estas cuatro decisiones son una teoría de toma de decisiones en sí misma, y cuentan con importantes restricciones: en la industria en general y en muchos otros ambientes. Los investigadores están inhibidos de usar combinaciones de factores que a priori se sepan que son peligrosas, que arruinen materia prima o se sepa que son ineficientes o que no funcionan.

Diseño factorial completo de 3^3	
000	-1-1-1
001	-1-1 0
002	-1-1 1
010	-1 0-1
011	-1 0 0
012	-1 0 1
020	-1 1-1
021	-1 1 0
022	-1 1 1
100	0-1-1
101	0-1 0
102	0-1 1
110	0 0-1
111 →centro←	0 0 0
112	0 0 1
120	0 1-1
121	0 1 0
122	0 1 1
200	1-1-1
201	1-1 0
202	1-1 1
210	1 0-1
211	1 0 0
212	1 0 1
220	1 1-1
221	1 1 0
222	1 1 1

Ya hemos mencionado los diseños factoriales completos. Un ejemplo es un diseño factorial completo de tres factores en tres niveles: bajo (**–1** o **0**), intermedio (**0** o **1**) y alto (**1** o **2**):

Este diseño en particular tiene un número de niveles idéntico para cada factor, pero pueden encontrarse diseños como:

2x3x4 para 2 niveles en el primer factor, 3 para el segundo y 4 para el último factor, que implica 24 ensayos,

2x3x3x4 de 2, 3, 3 y 4 niveles para los factores 1°, 2°, 3° y 4°. Que conlleva 72 ensayos.

En todo caso el número de ensayos sigue siendo

niveles$^{\mathbf{factores}}$ **= nivel de factor 1 x niveles de factor 2 x ... x niveles de factor f**

Box Wilson

Este método nos permite trabajar siempre cerca de condiciones de operación poco riesgosas o razonablemente cerca de un punto central que se supone suficientemente probado. (receta u operación de rutina, por ejemplo)

Las ventajas del método son:

a) varias mediciones o ensayos en el punto que se llamará central (000 en la tabla anterior, columna derecha),

b) puntos factoriales con solamente dos niveles (se excluye el punto central), generando un número de ensayos de $niveles^{factores} = 2^{factores}$

−1	−1
−1	+1
+1	−1
+1	+1

con tres factores,

−1	−1	−1
−1	−1	+1
−1	+1	−1
−1	+1	+1
+1	−1	−1
+1	−1	+1
+1	+1	−1

c) deberemos tomar la decisión sobre rotar o no el diseño. Con los datos obtenidos en el diseño hasta este punto el investigador puede determinar los componentes lineales de la superficie de respuesta, con lo cual podrá tomar la decisión de rotarlo o no. Con el conjunto de datos obtenidos en la rotación podrá analizar los componentes cuadráticos de la superficie de respuesta (curvatura). Esto supone ahorro de ensayos ya que no se realizan aquellos que se descubre que no son necesarios (decisión de no rotar)

d) puntos ortogonales. Para rotar el diseño se busca un factor de rotación K igual a la raíz cuadrada de los factores

Para el diseño de 2 factores aparecen los nuevos ensayos:

1,41	0
–1,41	0
0	–1,41
0	1,41

y con 3 factores:

–1,73	0	0
0	–1,73	0
0	0	–1,73
1,73	0	0
0	1,73	0

e) ajuste. Con los ensayos realizados se pueden ajustar los datos antes de rotar a una expresión:

$$z = b_0 + \sum_{j=1}^{m} b_j x_j$$

$$j = 1,2,..,m$$

si alguno de los b_j es diferente de cero no vale la pena continuar.

f) ajuste completo. Si los b_j son todos suficientemente parecidos a cero, se completa el ensayo mediante la rotación. El total de los datos, provenientes de antes y después de rotar, se ajusta a un modelo del tipo:

$$z = b_0 + b_1 x_1 + b_2 x_2 + b_{11} x_1^2 + b_{22} x_2^2 + b_{12} x_1 x_2$$ si m = 2

$$z = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_{11} x_1^2 + b_{22} x_2^2 + b_{33} x_3^2 + b_{12} x_1 x_2 + b_{13} x_1 x_3 + b_{23} x_2 x_3$$

si m=3

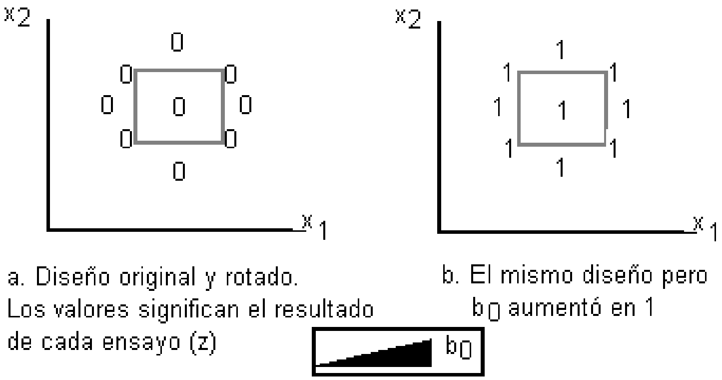
$$z = b_0 + \sum_{j=1}^{m} b_j x_j + \sum_{i=1}^{m} \sum_{j=1}^{m} b_{ij} x_i x_j$$ para cualquier m

g) identificación de los componentes del modelo[8]. Cada uno de estos elementos del modelo tiene significado, que es el siguiente:
b_0 es la altura promedio de la superficie de respuesta.

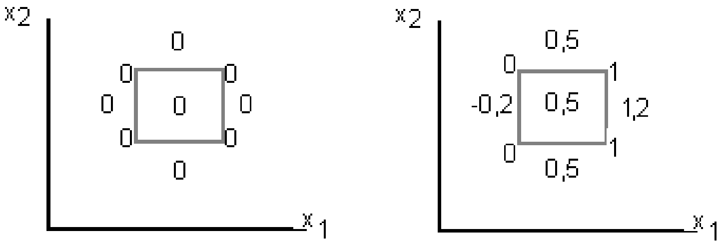
bi Cualquier valor distinto de cero significa que el factor *i* tiene influencia sobre los resultados. Un caso ejemplo sería que el factor diera un *Z* bajo en el nivel bajo, y alto en el alto: indica que, en esa dirección el funcional crece por influencia del factor. En el caso de las interacciones (curvatura), b_{ij} va a crecer hacia más de uno de los extremos respecto del punto central.

El significado de estos comportamientos son combinaciones o interrelaciones entre *i* y *j* favorables o no.

Ejemplo de variación de *bi* y *bij*:

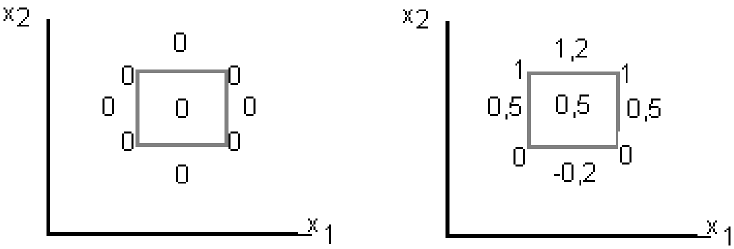


[8] Observamos que el modelo contiene todas las combinaciones lineales y cuadráticas posibles, pero no de orden superior



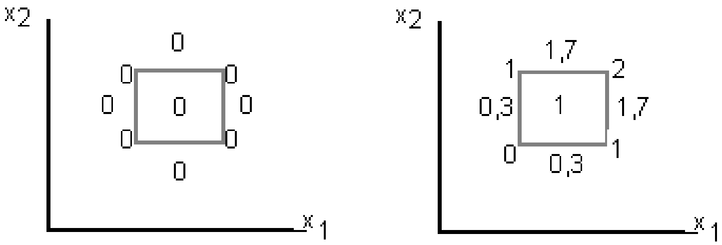
a. Diseño original y rotado.
Los valores significan el resultado
de cada ensayo (z)

b. El mismo diseño pero
b_1 aumentó



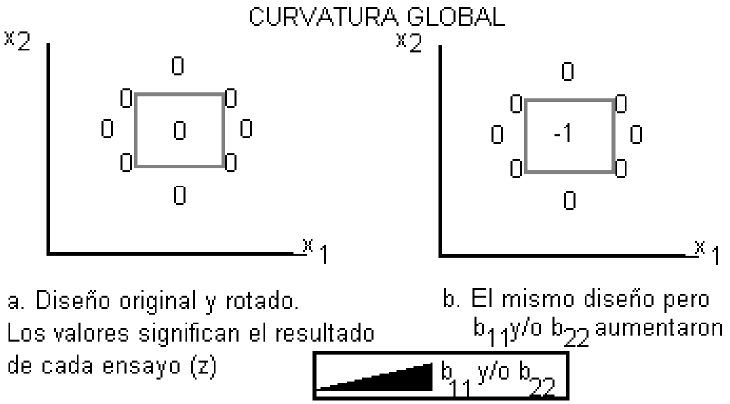
a. Diseño original y rotado.
Los valores significan el resultado
de cada ensayo (z)

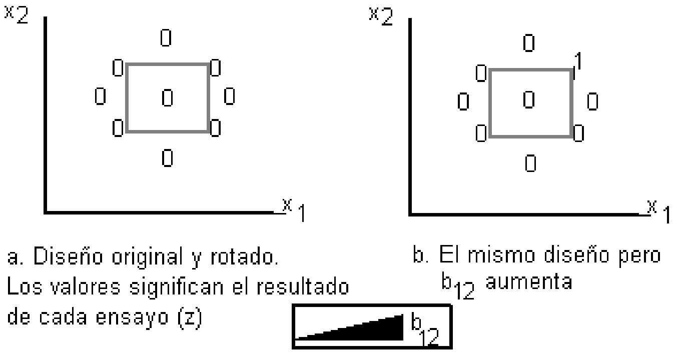
b. El mismo diseño pero
b_2 aumentó



a. Diseño original y rotado.
Los valores significan el resultado
de cada ensayo (z)

b. El mismo diseño pero
b_1 y b_2 aumentaron





Curvatura global

Resulta conveniente efectuar algunos comentarios respecto al concepto de curvatura global que vimos más arriba. En el modelo presentado, b_{11} y b_{22} miden curvaturas particulares, así b_{11} señala la importancia de la curvatura a lo largo del eje x_1 y b_{22} lo mismo a lo largo del otro eje. Si tenemos presente el modelo sobre el que estamos trabajando, que es:

$$z = b_0 + b_1 x_1 + b_2 x_2 + b_{11} x_1^2 + b_{22} x_2^2 + b_{12} x_1 x_2$$

podemos inferir que ni b_{11} tiene relación con el eje x_2 ni b_{22} tiene relación con el eje x_1, por lo cual nos damos cuenta que señalan curvaturas particulares.

Para poder establecer un modelo con curvatura global habría que modificar el anterior de la siguiente manera:

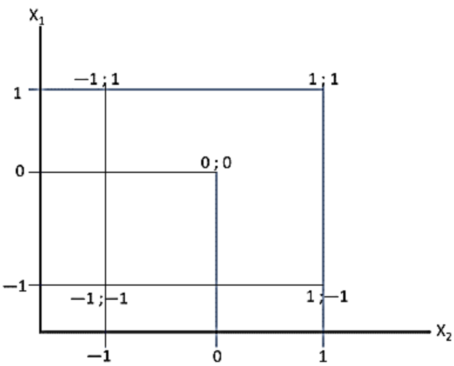
$$z = b_0 + b_1 x_1 + b_2 x_2 + b_{ii}(x_1^2 + x_2^2) + b_{12} x_1 x_2$$

allí aparece un nuevo coeficiente b_{ii} de curvatura global, que existe si cualquiera de las dos curvaturas particulares existe. Además, si existe (si no es nulo) no se sabe cuál de los dos ejes es el responsable, pudiendo ser, inclusive, ambos.

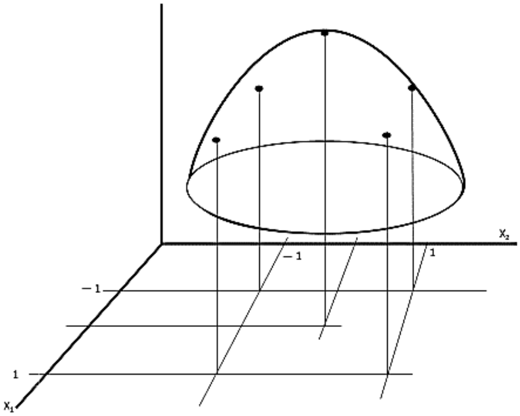
La importancia de este parámetro es que brinda información cuando se trabaja con pocos datos experimentales, es decir, con pocos puntos. En ese caso, la aparición de curvatura global, manifestada con b_{ii} diferente de cero, indica que en alguno de los ejes estamos cerca del óptimo. Esto significa que vale la pena hacer más experimentos para poder discriminar cuál o cuáles de los factores han provocado la curvatura. *Contrariu sensu*, un valor nulo en el coeficiente b_{ii} indica que las curvaturas particulares no tienen importancia.

Para poder trabajar con comodidad con la curvatura global se consideran los datos del punto central como datos de un nivel intermedio.

Podemos, entonces, definir la curvatura global como *la diferencia entre la respuesta del punto central y la hipotética respuesta que se obtendría si se ajustasen los puntos extremos a un modelo lineal* (o de primer orden).



El diseño en un plano x_1 — x_2 con dos factores y tres niveles (**—1**, **0** y **1**) se...



...puede ver en 3D imaginando una superficie de respuesta similar a media esfera

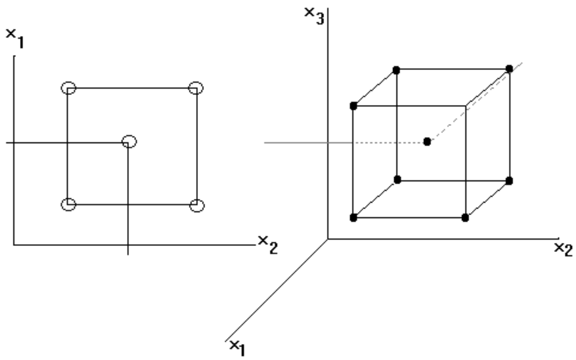
Estas figuras demuestran entonces que si se desea ajustar a un modelo simple son necesarios menos niveles que para un modelo de ajuste complejo. Cuando mayor es el orden del modelo, más niveles serán necesarios.

Tengamos presente que si un modelo de ajuste es complicado obtenemos una descripción más correcta de la superficie de respuesta determinada experimentalmente y menor el error que el mismo modelo no absorbe.

En otros términos: supongamos un modelo de tercer orden, conlleva — anidado — un modelo de segundo orden, así que es de esperar que va a interpretar los datos por lo menos tan bien como el de segundo orden y sobrarán algunos parámetros que permitirán describir algún comportamiento adicional de los datos experimentales. Desde ese punto de vista, el de tercer orden será mejor que el de segundo, pero la complejidad hace que el experimentador pueda perder el sentido físico de lo que se está buscando.

La curvatura global en un diseño

Desde el punto de vista de la investigación o el ensayo, ya sea en una industria o sobre un proceso, parece conveniente comenzar estimando curvatura del sistema entero. Si combinamos en un solo punto central del diseño a todos los factores que estamos estudiando, obtendremos diseños que podemos representar como las figuras de abajo para ensayos en un diseño 2^2 (izquierda) y de un diseño 2^3 (derecha)



La curvatura total de estos diseños se determina en forma similar que la curvatura simple, comparando el valor real experimental obtenido para el baricentro — coordenadas (0;0) — con el valor hipotético que se obtendría para ese punto si los puntos exteriores hubiesen estado ajustados a un modelo lineal: se compara con la media de los puntos 2^f

Puede ocurrir que
1) No se detecta curvatura. Entonces no conviene complicar las cosas: ajustando a un modelo lineal e interpretando los resultados se termina el ensayo
2) Se detecta curvatura. Entonces conviene planificar nuevos puntos experimentales distintos del central que permitan averiguar la incidencia de cada factor sobre la curvatura. Esos puntos se ubicarán, por ejemplo, siguiendo técnicas como la de Box Wilson.

Orden en la secuencia de ensayos

Los primeros ensayos serán los que nos permiten reconocer alguna estimación del error experimental. Por ejemplo, la desviación típica de una serie de ensayos hechos en igualdad de condiciones es una fuente para conocer el error.

Los siguientes pasos serán los que permitan encontrar el efecto de los factores más importantes y de la curvatura global sobre la superficie de respuesta. Esto lleva a un número reducido de ensayos y permite tener una idea de cuál es el camino más provechoso para avanzar en el diseño de experimentos.

Lo siguiente será dedicar esfuerzos en hallar el óptimo para aquellas variables de decisión que nos parecen que brindan la mayor expectativa o que creemos de que si se varían nos llevan a zonas deseables de la superficie de respuesta. Por último, buscaremos, en las variables, ajustes que no consideramos antes.

Ejemplo integrador

Tenemos un reactor *batch*, y hay acuerdo en adoptar un determinado criterio para definir el objetivo, ***Z***, que en este ejemplo es la eficiencia. Usando las especificaciones del vendedor realizamos dos corridas, con los siguientes resultados:

1º $\rightarrow$ Z = η = 50%
2º $\rightarrow$ Z = η = 52%

Como no se han alterado ninguna de las variables en esas corridas las usamos para calcular la varianza

$$s^2 = \frac{SumaCuadrados - \Pr omedio}{Grados.de.libertad} = \frac{50^2 + 52^2 - \frac{1}{2}(50+52)^2}{2-1} = 2$$

de donde podemos calcular directamente el error:

$$e = s = \sqrt{s^2} = 1{,}41$$

Nos proponemos estudiar, en lugar de todas las variables de decisión, solo aquellas que sean importantes y que signifiquen una mejora del 4,2% con respecto a la media hallada de 51. A ese valor, 4,2%, lo llamaremos Δ. Si disminuimos Δ debemos aumentar el número de ensayos.

Número de ensayos a realizar

Para estimar el número de ensayos, realizamos el cociente entre la variabilidad buscada y el error (Δ /e). Esta relación es importante: si vale 1 significa que la variabilidad por el efecto tiene la misma magnitud que la del error. Números mayores —por ejemplo 4— significan que podemos discriminar la variabilidad de los factores respecto a la que produce el error cuando aquella es 4 veces mayor. Números menores que 1 significan que la herramienta discrimina variabilidades pequeñas y enmascaradas por el error.

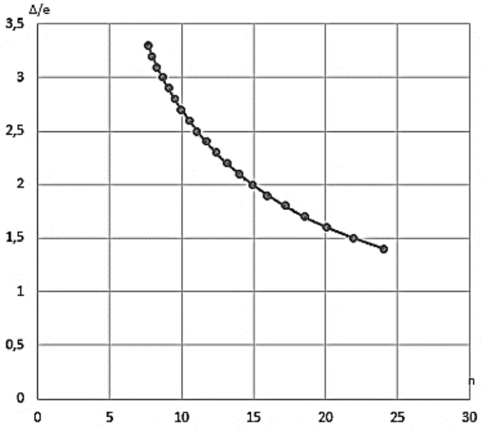
Es un buen punto de partida, para orientarnos, saber que se considera práctico realizar, a nivel industrial, un número máximo de 64 ensayos. Más ensayos no brindarían mejor información. (32 de cada tipo si son dos niveles y un factor, 16 de cada tipo si son dos niveles con dos factores, 8 de cada tipo si son dos niveles y tres factores, ...)[9]

Este número de ensayos se asocia a un Δ/e = 0,667: cuando el factor cambia de "bajo" a "alto" la respuesta cambia un 0,667 del total de cambios, inclusive los que produjo el error.

Podemos realizar un gráfico de la relación Δ/e y una tabla asociada

[9] No hay que confundir algunos términos. Con más ensayos puede haber mejor significado estadístico, pero en el caso de laboratorio o industria, escasos aportes prácticos. Pensemos que estamos investigando para encontrar la probabilidad de que aparezca un 6 arrojando un dado. Es evidente que luego de muchos ensayos, puede aparecer un sesgo en una de las caras del dado en particular que se usa, pero es muy difícil que si este sesgo es muy importante no se haya detectado a los pocos ensayos. *No quiere decir que la verdad* — según palabras del viejo maestro von der Becke — *sea poco valiosa en sí misma, sino que es poco útil en estos casos. Lo cual es diferente*.

Δ / e	número de ensayos necesarios
3,25	8
2,3	12
1,8	16
1,6	20
1,5	24



Para el caso en estudio, la relación [Δ /e] es [4,2/1,41] lo que significa que con 8 ensayos se podrá obtener la discriminación buscada y será posible estudiar un máximo de 7 variables. Si quisiéramos estudiar más aparecerán confusiones entre variables, confusiones originadas en un mal diseño de experimentos.

Proponemos un diseño con puntos factoriales y un punto central (las corridas realizadas, que se excluyen del diseño, no se cuentan entre los 8 ensayos primitivos). Los puntos factoriales se eligen de la siguiente manera:

$$\text{Nº ensayos} = \text{Niveles}^{\text{Factores}} = 2^7 = 128$$

este valor nos indica el tamaño, la cantidad de experimentos, de un diseño completo. Ese diseño completo se puede volcar en la **matriz de experimentos** que mostramos en la página siguiente.

En ella, la primera columna es x_1, la segunda x_2, y así hasta séptima, que es x_7. Para cada uno de los 8 bloques que representan todo el experimento.

Como este diseño es extravagantemente grande y complejo, lo vamos a transformar en uno incompleto de **8 ensayos**.

Para reducirlo de 128 a 8, o sea en un factor 16, (128/8 = 16) haremos cortes por la mitad. Como 16 = 2^4, significa que habremos de reducir en 4 cortes por la mitad:

0	0	0	0	0	0	0
0	0	0	0	0	0	1
0	0	0	0	0	1	0
0	0	0	0	0	1	1
0	0	0	0	1	0	0
0	0	0	0	1	0	1
0	0	0	0	1	1	0
0	0	0	0	1	1	1
0	0	0	1	0	0	0
0	0	0	1	0	0	1
0	0	0	1	0	1	0
0	0	0	1	0	1	1
0	0	0	1	1	0	0
0	0	0	1	1	0	1
0	0	0	1	1	1	0
0	0	0	1	1	1	1

0	0	1	0	0	0	0
0	0	1	0	0	0	1
0	0	1	0	0	1	0
0	0	1	0	0	1	1
0	0	1	0	1	0	0
0	0	1	0	1	0	1
0	0	1	0	1	1	0
0	0	1	0	1	1	1
0	0	1	1	0	0	0
0	0	1	1	0	0	1
0	0	1	1	0	1	0
0	0	1	1	0	1	1
0	0	1	1	1	0	0
0	0	1	1	1	0	1
0	0	1	1	1	1	0
0	0	1	1	1	1	1

0	1	0	0	0	0	0
0	1	0	0	0	0	1
0	1	0	0	0	1	0
0	1	0	0	0	1	1
0	1	0	0	1	0	0
0	1	0	0	1	0	1
0	1	0	0	1	1	0
0	1	0	0	1	1	1
0	1	0	1	0	0	0
0	1	0	1	0	0	1
0	1	0	1	0	1	0
0	1	0	1	0	1	1
0	1	0	1	1	0	0
0	1	0	1	1	0	1
0	1	0	1	1	1	0
0	1	0	1	1	1	1

0	1	1	0	0	0	0
0	1	1	0	0	0	1
0	1	1	0	0	1	0
0	1	1	0	0	1	1
0	1	1	0	1	0	0
0	1	1	0	1	0	1
0	1	1	0	1	1	0
0	1	1	0	1	1	1
0	1	1	1	0	0	0
0	1	1	1	0	0	1
0	1	1	1	0	1	0
0	1	1	1	0	1	1
0	1	1	1	1	0	0
0	1	1	1	1	0	1
0	1	1	1	1	1	0
0	1	1	1	1	1	1

1	0	0	0	0	0	0
1	0	0	0	0	0	1
1	0	0	0	0	1	0
1	0	0	0	0	1	1
1	0	0	0	1	0	0
1	0	0	0	1	0	1
1	0	0	0	1	1	0
1	0	0	0	1	1	1
1	0	0	1	0	0	0
1	0	0	1	0	0	1
1	0	0	1	0	1	0
1	0	0	1	0	1	1
1	0	0	1	1	0	0
1	0	0	1	1	0	1
1	0	0	1	1	1	0
1	0	0	1	1	1	1

1	0	1	0	0	0	0
1	0	1	0	0	0	1
1	0	1	0	0	1	0
1	0	1	0	0	1	1
1	0	1	0	1	0	0
1	0	1	0	1	0	1
1	0	1	0	1	1	0
1	0	1	0	1	1	1
1	0	1	1	0	0	0
1	0	1	1	0	0	1
1	0	1	1	0	1	0
1	0	1	1	0	1	1
1	0	1	1	1	0	0
1	0	1	1	1	0	1
1	0	1	1	1	1	0
1	0	1	1	1	1	1

1	1	0	0	0	0	0
1	1	0	0	0	0	1
1	1	0	0	0	1	0
1	1	0	0	0	1	1
1	1	0	0	1	0	0
1	1	0	0	1	0	1
1	1	0	0	1	1	0
1	1	0	0	1	1	1
1	1	0	1	0	0	0
1	1	0	1	0	0	1
1	1	0	1	0	1	0
1	1	0	1	0	1	1
1	1	0	1	1	0	0
1	1	0	1	1	0	1
1	1	0	1	1	1	0
1	1	0	1	1	1	1

1	1	1	0	0	0	0
1	1	1	0	0	0	1
1	1	1	0	0	1	0
1	1	1	0	0	1	1
1	1	1	0	1	0	0
1	1	1	0	1	0	1
1	1	1	0	1	1	0
1	1	1	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	1	0
1	1	1	1	0	1	1
1	1	1	1	1	0	0
1	1	1	1	1	0	1
1	1	1	1	1	1	0
1	1	1	1	1	1	1

Matriz de experimentos del diseño factorial completo

Un criterio de reducción podría ser eliminar aquellos ensayos en los que la suma de las columnas da impar 1(MOD 2) o par 0(MOD 2).

Si, por ejemplo, eliminamos las impares y dejamos las que cumplen con 0(MOD 2) para un grupo de variables:

$$x_1 + x_2 + x_3 + x_4 = 0(\text{MOD } 2)$$

Nos quedan 64 ensayos. A continuación, eliminamos las combinaciones que no cumplen con

$$x_3 + x_4 + x_5 + x_6 = 0(\text{MOD } 2)$$

Ahora nos quedan 32 ensayos. Luego dejamos las combinaciones que cumplan con

$$x_1 + x_3 + x_5 + x_7 = 0(\text{MOD } 2)$$

Tenemos 16 ensayos. Por último, dejamos las que cumplen con

$$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 = 1(\text{MOD } 2)$$

La tabla de los 8 ensayos remanentes será, entonces

0	0	0	0	1	1	1
0	0	1	1	0	0	1
0	1	0	1	0	1	0
0	1	1	0	1	0	0
1	0	0	1	1	0	0
1	0	1	0	0	1	0
1	1	0	0	0	0	1
1	1	1	1	1	1	1

No sería posible agregar una octava columna, ya que si partimos de $2^8 = 256$ ensayos veremos que al recortar hasta 8 nos quedarían dos columnas repetidas o dos columnas complementarias (con 1 donde la otra tiene 0), pero sí podemos agregar más ensayos centrales M, por ejemplo, M = 3.

El modelo de los 8 + 3 ensayos sería

$$z = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 + b_5 x_5 + b_6 x_6 + b_7 x_7 + b_8 (x_{21} + x_{22} + x_{23} + x_{24} + x_{25} + x_{26} + x_{27})$$

Lo importante es el último término. La función es algo así como una esfera de 7 dimensiones y el último término es lo que da la curvatura de esa esfera.

También debemos determinar los alias.

En primer lugar, reescribiremos los métodos de “corte” para fraccionar:

1) $x_1x_2x_3x_4$
2) $x_3 x_4 x_5 x_6$
3) $x_1 x_3 x_5 x_7$
4) $x_1 x_2 x_3 x_4 x_5 x_6 x_7$

Multiplicamos entre sí estas restricciones y omitimos los términos cuadráticos

1 por 2) $x_1 x_2 x_3 x_4 . x_3 x_4 x_5 x_6 = x_1 x_2 x_5 x_6$
1 por 3) $x_2 x_4 x_5 x_7$
1 por 4) $x_5 x_6 x_7$
2 por 3) $x_1 x_4 x_6 x_7$
2 por 4) $x_1 x_2 x_7$
3 por 4) $x_2 x_4 x_6$

Estas últimas expresiones las multiplicamos por las originales. Generan cinco restricciones más:

1 por2 por 3) $x_2 x_3 x_6 x_7$
1 por 2 por 4) $x_3 x_4 x_7$
1 por 3 por 4) $x_1 x_3 x_6$
2 por 3 por 4) $x_2 x_3 x_5$
1 por 2 por 3 por 4) $x_1 x_4 x_5$

Para averiguar los alias de x_1 multiplicaremos x_1 por cada relación y eliminaremos los términos cuadráticos. Los alias de x_1 serán:

$x_2 x_3 x_4$	$x_2 x_5 x_6$	$x_4 x_5$	$x_1 x_3 x_4 x_7$
$x_3 x_5 x_7$	$x_4 x_6 x_7$	$x_1 x_2 x_3 x_5$	$x_1 x_2 x_3 x_6 x_7$
$x_2 x_7$	$x_2 x_7$	$x_1 x_2 x_5 x_6$	$x_1 x_2 x_4 x_5 x_7$
$x_2 x_3 x_4 x_5 x_6 x_7$	$x_3 x_6$	$x_1 x_5 x_6 x_7$	$x_1 x_3 x_4 x_5 x_6$

Si suponemos que decidimos considerar con significado físico a solo cuatro de las ocho variables y a la curvatura (x_1, x_2, x_3, x_4 y x_8) quiere decir que las variables restantes (x_5, x_6 y x_7), no tendrán significado propio, serán variables *slacks* o flojas. Sin embargo, los alias de estas variables pueden tener significado físico. Por ejemplo, la interacción entre x_4 y x_5 —una real y una floja— deja de tener significado físico, pero en cambio aporta como estimación del error. ¿Cuál es el objeto de considerar importante estas variables?

Si la variable X_5, en el funcional aparece como $\mathbf{b_5} \boldsymbol{X_5}$., podremos replantear examinando los alias de X_5 y eliminando aquellos alias que conlleven variables flojas, como:

$$b_5 x_5 = b_5 (x_1 x_4 + x_2 x_3)$$

Ahora estamos en condiciones de atribuir significado físico a las cuatro variables que arbitrariamente hemos seleccionado. Vamos a distribuir esas variables, y sus niveles, (llamando nivel 0 al punto central) de la siguiente manera:

FACTOR	NIVEL —1	NIVEL 0	NIVEL 1
Temperatura (ºC)	70	76	80
Exceso estequiométrico	5%	10%	15%
Duración bache (min)	120	150	180
Concentrac. catalizador	1%	2%	3%

Podemos simular los datos correspondientes a los puntos experimentales. Para eso vamos a usar un modelo como el que sigue:

$$z = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \varepsilon_m + e_{ijklmn}$$

donde:

μ verdadera media de los 11 ensayos. Puede ser cero si ese valor es 51

α_i es la incidencia del factor x_1 (temperatura). Vale 3 a nivel —1, 0 a nivel 0 y —3 a nivel 1.

β_j idem, factor x_2, exceso. 1 en nivel —1, 0 en nivel 0 y —1 en nivel 1

γ_k idem, factor x_3, duración. 5 en nivel —1, 0 en nivel 0 y —5 en nivel 1

δ_l idem, factor x_4, concentración catalítica, —3 en nivel —1, 0 en nivel 0 y 3 en nivel 1

ε_m idem, factor x_8, curvatura, que se establece en cero para todos los niveles

e_{ijklmn} incidencia aditiva del error experimental. Se toman por ejemplo los siguientes valores para cada uno de los ensayos:

$$3, 0, -1, -1, 1, 1, -2, 2, 0, 1, 0, \text{ y } -3.$$

Con esa expresión, entonces, "fabricamos" los "datos experimentales", supongamos que son los que tabulamos en la tabla siguiente:

Ensayo	Punto	Función	Dato
1	0000111	0+3+1–5–3+0+3	–1
2	0011001	0+3+1+5+3+0+0	12
3	0101010	0+3–1–5+3+0–1	–1
4	0110100	0+3–1+5–3+0–1	3
5	1001100	0–3+1–5+3+0+1	–3
6	1010010	0–3+1+5–3+0–2	–2
7	1100001	0–3–1–5–3+0+2	–10
8	1111111	0–3–1+5+3+0+0	4
9	0000000	0+0+0+0+0+0+1	1
10	0000000	0+0+0+0+0+0+0	0
11	0000000	0+0+0+0+0+0–3	–3
Suma			0

Podemos usar hoja de cálculo para ingresar todo el problema, como vemos en la siguiente tabla.

Como usaremos la facilidad "Análisis de Regresión" de la hoja de cálculo Excel, entonces nos adaptaremos a la terminología que se utiliza en ella.

Así, la columna Z será a partir de ahora el ***Rango "Y"*** y la matriz [x1:x8] será el ***Rango "X"***:

f	x1	x2	x3	x4	x5	x6	x7	x8	Z
1	–1	–1	–1	–1	1	1	1	0,375	–1
1	–1	–1	1	1	–1	–1	1	0,375	12
1	–1	1	–1	1	–1	1	–1	0,375	–1
1	–1	1	1	–1	1	–1	–1	0,375	3
1	1	–1	–1	1	1	–1	–1	0,375	–3
1	1	–1	1	–1	–1	1	–1	0,375	–2
1	1	1	–1	–1	–1	–1	1	0,375	–10
1	1	1	1	1	1	1	1	0,375	4
1	0	0	0	0	0	0	0	–1	1
1	0	0	0	0	0	0	0	–1	0
1	0	0	0	0	0	0	0	–1	–3
∑= 11	0	0	0	0	0	0	0	0	0

Para la variable x8 asignamos un valor positivo en los puntos factoriales y uno negativo en los centrales. El efecto de curvatura queda así neutralizado entre todos los valores.

Si ahora hacemos el ***análisis de regresión X—Y***, obtendremos las tablas que siguen.

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0,98515052
Coeficiente de determinación R^2	0,97052154
R^2 ajustado	0,85260771
Error típico	2,081666
Observaciones	11

ANÁLISIS DE VARIANZA

	Grados de libertad	Suma de cuadrados	Promedio de cuadrados	F	Valor crítico de F
Regresión	8	285,333333	35,66666667	8,23076923	0,11280166
Residuos	2	8,66666667	4,333333333		
Total	10	294			

	Coef.	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%	Inferior 95.000%	Superior 95.000%
Intercepción	0	0,62764591	0	1	-2,70054429	2,70054429	-2,70054429	2,70054429
x1	-3	0,73598007	-4,07619732	0,05524501	-6,16666887	0,16666887	-6,16666887	0,16666887
x2	-1,25	0,73598007	-1,69841551	0,23152672	-4,41666887	1,91666887	-4,41666887	1,91666887
x3	4	0,73598007	5,43492976	0,03222659	0,83333113	7,16666887	0,83333113	7,16666887
x4	2,75	0,73598007	3,73651421	0,06474667	-0,41666887	5,91666887	-0,41666887	5,91666887
x5	0,5	0,73598007	0,67936622	0,5669873	-2,66666887	3,66666887	-2,66666887	3,66666887
x6	-0,25	0,73598007	-0,33968311	0,76645032	-3,41666887	2,91666887	-3,41666887	2,91666887
x7	1	0,73598007	1,35873244	0,30717968	-2,16666887	4,16666887	-2,16666887	4,16666887
x8	0,66666667	1,02494149	0,65044363	0,58214455	-3,74330369	5,07663702	-3,74330369	5,07663702

Como vemos, existe información dispersa sobre todo el modelo, sin restricciones, aunque un bajo grado de significación en el F correspondiente a la regresión.

Si reelaboramos el proceso solamente con las variables 1, 2, 3 y 4 y la 8, encontraremos resultados más concretos, como los que aparecen en las tablas que siguen.

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0,96685431
Coeficiente de determinación R^2	0,93480726
R^2 ajustado	0,86961451
Error típico	1,95789002
Observaciones	11

ANÁLISIS DE VARIANZA

	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	F	Valor crítico de F
Regresión	5	274,833333	54,9666667	14,3391304	0,00548867
Residuos	5	19,1666667	3,83333333		
Total	10	294			

(vemos que ahora F es significativo)

	Coeficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%	Inferior 95.000%	Superior 95.000%
Intercepción	0	0,59032605	0	1	-1,51747895	1,51747895	-1,51747895	1,51747895
x1	-3	0,69221866	-4,33389071	0,00747186	-4,77940179	-1,22059821	-4,77940179	-1,22059821
x2	-1,25	0,69221866	-1,8057878	0,13077954	-3,02940179	0,52940179	-3,02940179	0,52940179
x3	4	0,69221866	5,77852095	0,0021831	2,22059821	5,77940179	2,22059821	5,77940179
x4	2,75	0,69221866	3,97273315	0,01060724	0,97059821	4,52940179	0,97059821	4,52940179
x8	0,66666667	0,96399841	0,69156407	0,52000281	-1,81136608	3,14469941	-1,81136608	3,14469941

Esta última tabla indicaría que el modelo de comportamiento es:

$$Z = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 + b_8 x_8$$

reemplazando

$$z = 0 - 3x_1 - 1{,}25\, x_2 + 4\, x_3 + 2{,}75\, x_4 + 0{,}67\, x_8$$

que difiere en muy poco (en realidad en el error) del modelo utilizado para generar los datos que, si se reemplazan por el coeficiente que corresponde sería

$$z = 0 - 3\, x_1 - 1\, x_2 + 5\, x_3 + 3\, x_4 + 0\, x_8$$

Queda solo preguntarse, ¿Como podemos averiguar el grado de significación de cada una de las variables e interrelaciones que representan? ¿Fue correcto haberlas excluido?

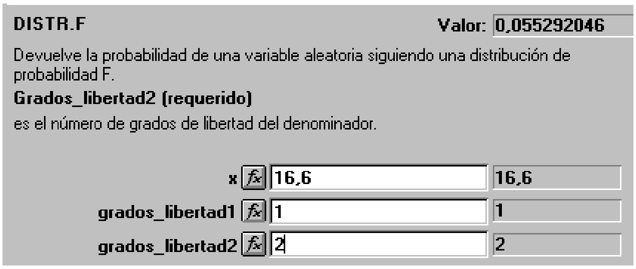
Volvemos a la tabla de la primera corrida con todas las variables.

	Coeficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%	Inferior 95.000%	Superior 95.000%
Intercepción	0	0,62764591	0	1	-2,70054429	2,70054429	-2,70054429	2,70054429
x1	-3	0,73598007	-4,07619732	0,05524501	-6,16666887	0,16666887	-6,16666887	0,16666887
x2	-1,25	0,73598007	-1,69841555	0,23152672	-4,41666887	1,91666887	-4,41666887	1,91666887
x3	4	0,73598007	5,43492976	0,03222659	0,83333113	7,16666887	0,83333113	7,16666887
x4	2,75	0,73598007	3,73651421	0,06474667	-0,41666887	5,91666887	-0,41666887	5,91666887
x5	0,5	0,73598007	0,67936622	0,5669873	-2,66666887	3,66666887	-2,66666887	3,66666887
x6	-0,25	0,73598007	-0,33968311	0,76645032	-3,41666887	2,91666887	-3,41666887	2,91666887
x7	1	0,73598007	1,35873244	0,30717968	-2,16666887	4,16666887	-2,16666887	4,16666887
x8	0,66666667	1,02494149	0,65044363	0,58214455	-3,74330369	5,07663702	-3,74330369	5,07663702

y la transformamos en una tabla de análisis de varianza:

	Coeficientes	SC Suma cuadr	GL	CM	F
Intercepción	0				
x1	–3	72	1	72	16,6
x2	–1,25	12,5	1	12,5	2,9
x3	4	128	1	128	29,6
x4	2,75	60,5	1	60,5	14,0
x5	0,5	2	1	2	0,5
x6	–0,25	0,5	1	0,5	0,1
x7	1	8	1	8	1,8
x8	0,6666666	1,84	1	1,84	0,4
Suma (regresión)		285,34	8	35,7	8,2
Residual por diferencia		8,66	2	4,33	
Total		294	10		

Para cada valor F (obtenido mediante el cociente de los cuadrados medios de cada variable respecto a los cuadrados medios del error), podemos hacer un análisis con la función DISTR.F, que nos devuelve la probabilidad.



El criterio que emplearemos es:

Muy significativo	p ≤ 0,001
Significativo	p ≤ 0,01
Moderadamente Significativo	p ≤ 0,05

En este primer valor, p = 0,055, lo que nos indica que el factor temperatura es significativo.

Para las otras variables obtenemos las siguientes probabilidades:

x_1	16,6	0,05529205	No Significativo
x_2	2,9	0,23069074	No significativo
x_3	29,6	0,03216279	Moderadamente Significativo
x_4	14	0,06458565	Significativo
x_5	0,5	0,5527864	No significativo
x_6	0,1	0,78178211	No significativo
x_7	1,8	0,3117528	No significativo
x_8	0,4	0,59175171	No significativo
r	8,2	0,10338327	No significativo

Ahora estamos en condiciones de reconstruir la tabla, asignando más grados de libertad al residual (error) y eliminando las variables no usadas por ser no significativas:

	Coeficientes	SC Suma cuadr	GL	CM	F
Intercepción	0				
x1	–3	72	1	72	18,7
x2	–1,25	12,5	1	12,5	3,2
x3	4	128	1	128	33,3
x4	2,75	60,5	1	60,5	15,75
x8	0,66666667	1,84	1	1,84	0,5
Suma (regresión)		274,8	5	55	14,3
Residual por diferencia		19,2	5	3,84	
Total		294	10		

x1	18,7	0,00298397	Significativo
x2	3,2	0,1137739	No significativo
x3	33,3	0,00076394	Muy significativo
x4	15,8	0,00443201	Poco significativo
x8	0,5	0,76748868	No significativo
r	14,3	0,0055228	Significativo

Podemos concluir en que la influencia del tiempo en la eficiencia, $x_{3,}$ es dominante, sigue la temperatura, pero con signo negativo (hay que disminuirla), luego x_4 y por último $x_{2,}$ el exceso. ¿Cuál será el ensayo que tenga en cuenta estos resultados?

Hay varias técnicas, una es fijar un “paso” como búsqueda simplex, por ejemplo, 0,5 y averiguar el valor de cada $x_{i:}$

general = coeficiente del factor dominante / paso = (4) / (0,5) = 8

x_1 ➔ paso = (Coeficiente)/(general) = (– 3) / (8) = – 0,3775
x_2 ➔ paso = –1,25 / 8 = – 0,15
x_3 ➔ paso = 4 / 8 = 0,5
x_4 ➔ paso = 0,67 / 8 = 0,084

El nuevo punto se realizará, entonces en las coordenadas

$$x_1 = -0,375;\ x_2 = -0,15;\ x_3 = 0,5 \text{ y } x_4 = 0,084.$$

Una vez obtenido el valor del funcional iremos determinando la variación de ese valor respecto a los anteriores.

Podemos hacer así una búsqueda directa simplemente agregando un nuevo ensayo en la dirección de la pendiente más significativa y a una “distancia” de un paso.

Nota importante:

El ejemplo que acabamos de desarrollar parte de una premisa equivocada en la elección del criterio: si se busca un rendimiento máximo en una superficie aparentemente lineal (excepto la temperatura que está asociada a la cinética de la ecuación) y sin restricciones, es probable que los tiempos tiendan a infinito. Sin embargo, lo hemos incluido y utilizado porque tiene valor didáctico y con ese fin, exclusivamente. No pretende ser demostrativo de una situación estrictamente real.

CAPÍTULO 21
Software para diseño de experimentos.

Como hemos visto en los diferentes libros de Optimiza 12 para cada tema hicimos referencias al software adecuado disponible, con particular preferencia hacia las opciones gratuitas, o, al menos, de software libre, con la excepción de WinQSB, que, si bien en sus orígenes no fue gratuito, la obsolescencia y fácil disponibilidad como *freeware* lo han convertido en una opción ineludible para tener en cuenta en la fase de aprendizaje de algunos temas, aunque con poca utilidad en cuanto a posibilidades de aplicación profesional.

La otra excepción es ***Excel***, que, si bien no es gratuito, es de uso muy común y extendido en cualquier ambiente profesional y, en muchas de sus funciones, existen versiones alternativas de código abierto. Lo mismo podemos decir respecto a ***Project*** y sus alternativas.

No es el caso de del software disponible para diseño de experimentos o, genéricamente, para análisis estadístico de datos.

En este caso, en los capítulos precedentes hacemos referencia a las capacidades de ***Excel***, y de las hojas de cálculo en general, para ordenar datos y hacer análisis estadístico, pero no hemos hecho referencia a los productos destinados especialmente a este fin.

Las razones son varias, y las iremos desarrollando, pero el componente fundamente es el elevado costo de las licencias de la mayoría de los programas destinados al análisis estadístico de datos.

La que sigue es una brevísima reseña de algunas de las herramientas que hemos analizado y que se pueden destinar al diseño de experimentos. Dejamos sentado que existen muchas más que las que mencionamos y que, de ninguna manera, hacemos una clasificación valorativa de las cualidades, ya que ellas varían según el destino de su uso y las habilidades personales de cada usuario.

Podríamos hacer clasificaciones por tipo de análisis o complejidad de uso o alcance de la utilización, pero solo haremos una simple clasificación de lo que consideramos recomendable para tipo de necesidad que la práctica profesional nos presente, repetimos, siempre dentro de un universo acotado de programas. Sería imposible evaluar todo lo disponible.

R

El inicio de este recorrido necesariamente debe comenzar con "***R***". Se trata de un ***lenguaje de programación*** específicamente diseñado para estadística y que, para usarlo correctamente, necesitaremos dedicar una buena cantidad de tiempo al aprendizaje. Por eso sus usuarios habituales son especialistas en estadística o matemática avanzada que se han dedicado a dominar ***R*** y hacer sus propios desarrollos.

En la bibliografía de este Libro citamos un trabajo de Aranda de la Universidad de Sevilla que muestra aplicaciones ***R*** para diseño de experimentos, en el marco de una tesis.

Los siguientes programas que mencionamos a partir de acá son más sencillos de usar y de aprender, aunque con ciertas diferencias, algunas de ellas muy importantes. Por otra parte, la mayor parte de los programas "clásicos" que se usan en estadística (*Statgraph*, por citar uno) tienen complementos o posibilidades de usarlos en relación con ***R***, por lo cual nos parece necesaria esta mención inicial para que sepamos de que se trata y cuáles son sus alcances. Inclusive, en la reseña que sigue, vamos a encontrarnos con otras aplicaciones relacionadas con este lenguaje.

Hojas de cálculo. Excel y Calc

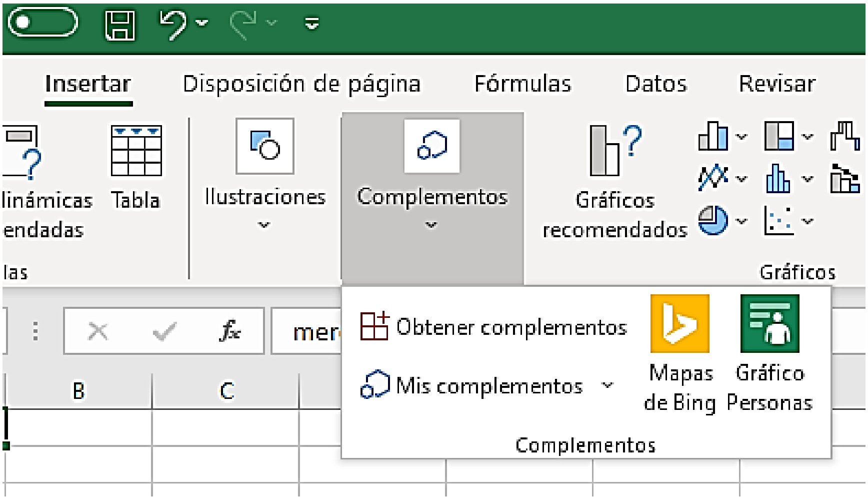
Ya mencionamos a ***Excel***. La ventaja enorme de este programa y sus opciones de software libre radica en que ofrece la posibilidad de programar sus macros mediante *VisualBasic for Aplications*, lo cual nos abre, como usuarios, un abanico de enormes alternativas, sin la magnitud de ***R*** pero mucho más simple para aprender y utilizar.

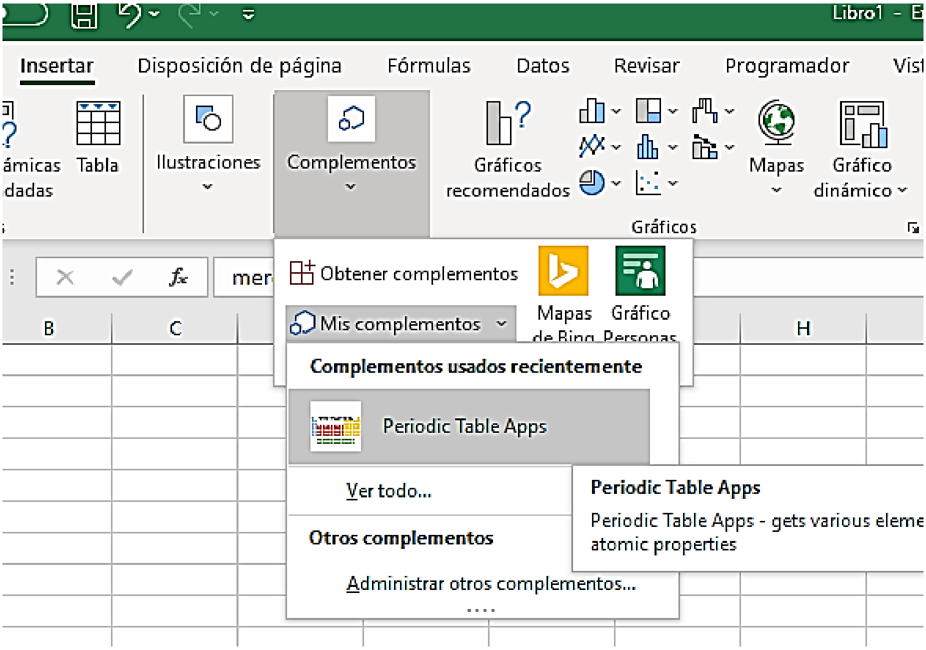
Además, tenemos la oferta de los complementos ya diseñados para usar en Excel para este y otros fines. En este caso hemos encontrado uno de la editorial *Prentice Hall*, llamado ***PH Stat***, que aparentemente fue retirado del sitio web de la editorial, pero que aún se podría conseguir si hemos adquirido ciertas publicaciones de libros de la editorial. Además, *Prentice-Hall/Pearson* ofrece una nueva herramienta informática que aparentemente es el reemplazo de ***PH Stat***, que ahora funciona *on line,* se denomina ***Statcrunch***, y se encuentra en la página de *Prentice*, aunque el acceso está limitado a la adquisición de textos y/o consecuente suscripción. (www.statcrunch.com) (www.pearson.com).

En cuanto a los ***complementos de Excel***, el que destaca es uno que nos permite que conectemos nuestra hoja y sus datos y procesos estadísticos con el programa ***Power BI*** (del cual hablaremos un poco más adelante). Este complemento tiene una versión gratuita para usuarios ***Excel***.

Otro complemento que nos puede resultar útil, aunque no es gratuito, es "***Surface chart***" que permite dibujar superficies de respuesta en 3D. Tiene una versión de prueba gratuita.

Para acceder a los ***complementos de Excel*** tendremos que buscar la pestaña "*Insertar*", luego ir al desplegable "*Complementos*" donde nos encontraremos con varias opciones: obtener complementos o acceder a una lista de los instalados, para poder usarlos, como vemos en la figura siguiente:





Nuestra recomendación es usar lo más posible la hoja de cálculo, dado su relativo bajo costo y disponibilidad.

Programas especializados.

Hasta ahora hemos mencionado programas generales, el más simple (***Excel***) o el más complicados (***R***) que, en realidad es un programa para "fabricar" programas o herramientas.

Haremos ahora un recorrido muy superficial por algunos programas que fueron diseñados específicamente para tareas estadísticas de análisis de grandes volúmenes de datos o de resultados de experimentos.

Power BI

Si bien es un producto de *Microsoft*, de precio modesto (hasta hace poco era gratis), no pertenece a la familia "***Office***" y tiene aspectos dedicados a esta disciplina que lo convierten en una herramienta especializada más avanzada que ***Excel*** (en lo que se refiere al campo de estadística de datos) con la posibilidad de acceder a los datos[10] que producen otras muchas aplicaciones en tiempo real.

Entre sus capacidades está la de manejar una cantidad de datos mayor que ***Excel*** y la de producir información sencilla y flexible. La contrapartida es que no posee incorporados cálculos estadísticos, eso hace que hay que usar otro software para hacer análisis estadístico, como ya dijimos ***Excel*** (para lo cual viene bien el complemento que mencionamos más arriba), o ***R*** o ***Python***[11]. La ventaja mayor es que es excelente y que ofrece un abono mensual muy accesible y es muy fácil de usar.

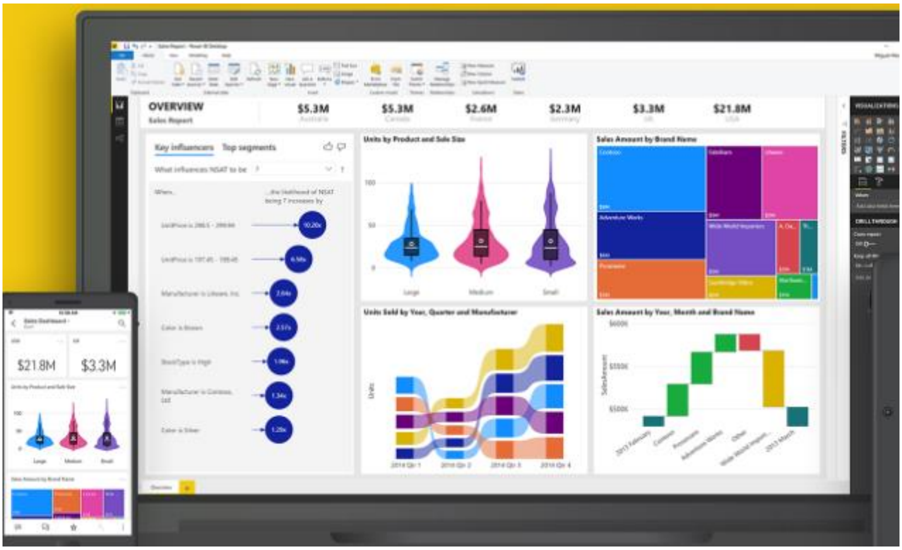
En el sitio oficial de ***Power Bi*** (https://powerbi.microsoft.com/es-es/.) hay tutoriales, videos, ejemplos y cursos para su uso y se disponen de **versiones gratuitas**.

---

[10] Puede llamarnos la atención que haya tanta disponibilidad (y altos precios) de programas capaces de trabajar estadísticamente con datos y hasta en tiempo real. Esto deriva del fenómeno que se denomina "*big data*" (macrodatos, datos masivos, inteligencia de datos, etc.) y que deriva de la aplicación de tecnologías de la información y comunicación (llamadas *TIC*) que permiten analizar en tiempo real patrones de comportamiento en aquello que se relaciona con casi cualquier actividad humana o de robots.

Así se genera la necesidad de disponer de herramientas capaces de recolectar, almacenar, buscar, compartir, analizar y visualizar datos para que ellos puedan ser de alguna utilidad para alguien. Se da la paradoja que tener enormes cantidades de datos puede llegar a ser tan poco útil como no tener datos en absoluto.

[11] ***Python*** es otro lenguaje de programación de los varios que existen en ese nivel. (C++, Pearl, Java Script, C#, PHP, etc. etc.)



SPSS Stadistics

Probablemente sea el mejor software disponible. Es muy simple y brinda enormes posibilidades, por lo cual es muy utilizado en los ambientes relacionados con la ingeniería, biología, química, estudios de mercado, salud, etc.

El inconveniente es la licencia necesaria para usarlo que tiene costo elevado. ***IBM*** de Argentina ofrece una versión de prueba gratuita y una cotización bajo solicitud para estudiantes en el sitio web https://www.ibm.com/ar-es/products/spss-statistics/pricing

Las instrucciones para el acceso de prueba para estudiantes a la versión *on line* por suscripción se encuentran en el sitio web de IBM, en el siguiente enlace: https://www.ibm.com/ar-es/products/spss-statistics/details. Su uso es sencillo y bastante intuitivo.

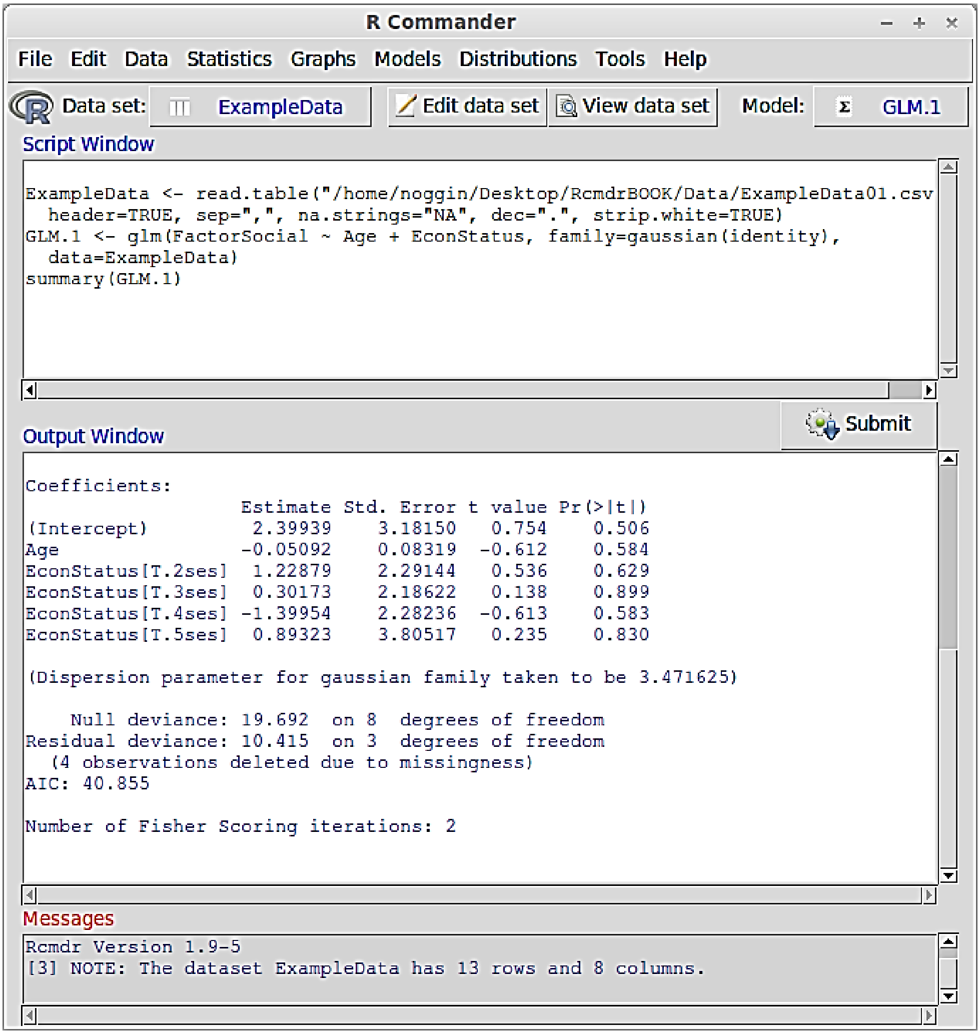
RCommander

Para aquellos que no pueden acceder a ***SPSS*** por el precio o a ***R*** por las dificultades para usarlo, existe una opción **gratuita** que es muy recomendable llamada ***RCommander***.

Como ya dijimos, ***R*** es el lenguaje de estadística por excelencia. De hecho, el lenguaje ***R*** es aplicable a la ciencia de datos y su manejo es complejo, ya que es necesario saber programar y la curva de aprendizaje es muy empinada.

Para los que no quieran programar, ***RCommander*** es una alternativa accesible. Se trata de una interfaz de usuario (*framework*) que ayuda a utilizar con sencillez muchas de las funcionalidades de ***R***. Es adaptable a diversas opciones y hasta se pueden personalizar los cálculos. Sin embargo, no es tan amigable e intuitivo como ***SPSS***. Las ventajas son que es gratuito y que incorpora y permite acceder a un lenguaje libre muy potente, ***R***. Es la opción indiscutida a ***SPSS*** para quien no quiera (o no pueda) gastar en la licencia necesaria para usarlo.

La página oficial del programa (www.rcommander.com) ofrece la descarga del producto, complementos y tutoriales.



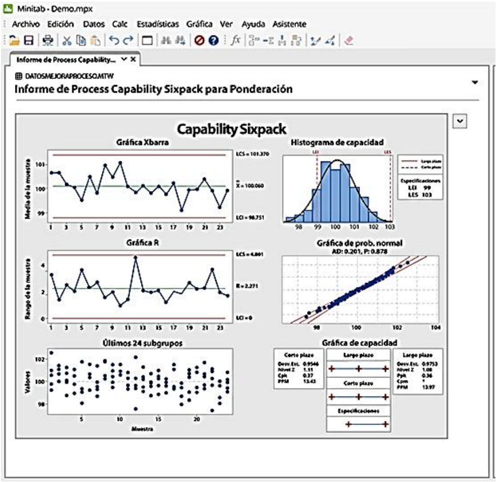
MiniTAB

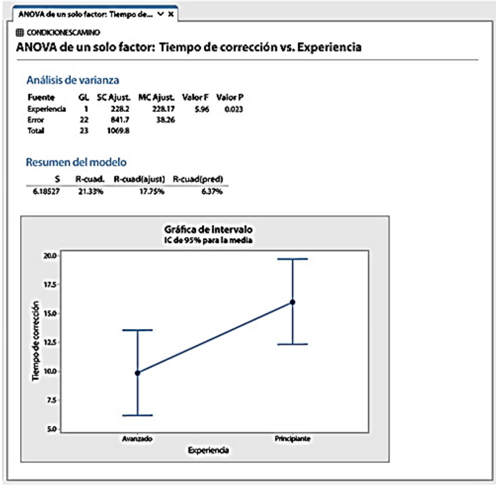
NUEVA VERSIÓN

Minitab 19

Minitab 19 es un software que tiene apariencia similar a ***SPSS***, y no solo la apariencia, también el precio es bastante elevado. En la página oficial del programa se ofrecen versiones de prueba que solo pueden utilizarse por un período breve de tiempo. La curva de aprendizaje es simple y con grandes posibilidades de uso, aunque no alcanza la magnitud de ***SPSS***.

El sitio web oficial es www.minitab.com/es-mx y allí hay una versión de prueba disponible por un mes.





InfoStat

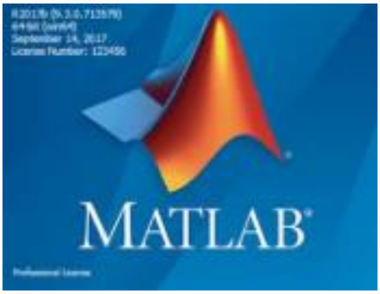
Se trata de un software para análisis estadístico de aplicación general desarrollado bajo la plataforma Windows por docentes de la *Universidad Nacional de Córdoba.* Está disponible para uso por estudiantes en la *Universidad Nacional de Luján* y además se pueden encontrar versiones gratuitas tanto para estudiantes como de prueba en la página www.infostat.com.ar.

Cubre tanto las necesidades elementales para la obtención de estadísticas descriptivas y gráficos para el análisis exploratorio, como métodos avanzados de modelación estadística y análisis multivariado. Una de sus fortalezas es la sencillez de su interfaz combinada con capacidades profesionales para el análisis estadístico y el manejo de datos. Debido al origen universitario, el programa tiene muchas facilidades para la enseñanza de la estadística que no son fáciles en encontrar en otros programas similares. Se dispone en versiones en varios idiomas, entre ellos, por supuesto, en español que aporta más ventajas desde el punto de vista de los estudiantes usuarios.

Una propiedad casi única entre el software estadístico es la habilidad de ***InfoStat*** de vincularse con ***R***, de dos maneras:

- mediante un intérprete integrado que permite ejecutar código escrito en ***R*** (script) sin salir del ambiente de trabajo de ***InfoStat*** y
- mediante la posibilidad de desarrollo de aplicaciones utilizando el motor de cálculo de ***R*** pero con una interfaz amigable para los los usuarios y que ***InfoStat*** ofrece para modelos lineales mixtos y generalizados, los cuales siempre fueron difíciles de especificar por su complejidad. De esta forma se pueden incorporar esos contenidos en cursos y capacitaciones que de otra manera hubieran sido de abordaje complejo.

MatLab



Matlab es un clásico de los paquetes de matemáticas, que debe su éxito a la cantidad de complementos que le agregan funcionalidades y a su potencia en muchos campos.

Se está extendiendo su uso en grandes empresas para desarrollo de aplicaciones en ingeniería. Es especialmente bueno en dinámica de sistemas, y un poco menos potente en estadística.

Se pueden descargar versiones de prueba en
https://la.mathworks.com/products/matlab.html?s_tid=hp_products_matlab

Tiene la capacidad de transformar las salidas en lenguaje ***C***+ para uso en dispositivos embebidos y dispone de una cantidad de módulos documentados y probados.

Frente a la desventaja que representa una complicada curva de aprendizaje encontramos la ventaja de que su costo es muchísimo menor que ***MiniTAB*** o ***SPSS***, aun incluyendo los módulos de programación lineal y de estadísticas avanzadas., lo cual se suma a que es posible que lo encontremos en empresas de ingeniería.

RStudio

En realidad, la definición correcta de ***RStudio*** sería la de un “ambiente” de trabajo (*framework*) que nos permite utilizar ***R*** desde una interfaz de usuario de buena calidad.

Además de las facilidades de uso para programar en ***R***, de forma sencilla nos permite usar e incorporar a nuestro proyecto paquetes de funciones ya hechos.

Las posibilidades de ***RStudio*** son amplias, porque, entre otras cosas, podríamos crear programas usando ***R*** sin necesidad de especializarnos en el lenguaje en sí, pero además podemos hacer muy buenos informes rápidos, presentación de análisis estadístico de datos y realizar gráficos muy potentes y profesionales.
A las características anteriores se le suma que se trata de un buen producto gratuito y fácil de adquirir. Se lo encuentra en la dirección web www.rstudio.com.

Python y otros lenguajes de programación

Es un excelente lenguaje de programación, que, según muchas opiniones expertas, es una poderosa herramienta que debería ser adoptada sin dudas por quienes quieren dedicarse a programar.

La complejidad del uso y del aprendizaje es mayor que la que se tiene con ***R*** ya que solo es posible usarlo para programar y eso es complejo pero potente. Es gratuito, hay una biblioteca completa de información y de ambientes de trabajo para usarlo. Se recomienda para estudiantes de disciplinas relacionadas con la informática y hemos visto resultados brillantes obtenidos por alumnos en nuestra práctica docente.

Como ya mencionamos más arriba, en esta categoría de lenguajes multiplataforma, podemos mencionar a ***Python*** y ***R***, pero existen alternativas diversas. Quizá es tan importante como el propio lenguaje sea el ambiente que elijamos para cada uno de ellos, como ya mencionamos para ***R***. Hay ambientes disponibles, casi todos gratuitos, para **C**, ***C+***, ***C#***, ***JavaScript***, ***VisualBasic***, ***Pearl***, ***PHP*** etc. y algunos de ellos sirven para más de un lenguaje. Como ejemplo, Microsoft ofrece ***Visual Studio***, que es de código abierto y gratuito y que soporta buena parte de los lenguajes, con la ventaja de ofrecer “librerías”, soporte, tutoriales y complementos. Otros editores con buenas prestaciones y facilidades son ***Atom***, ***Sublime***, ***Notepad***+, ***Eclipse***, y más. En general son de código abierto y gratuitos.

Conclusiones

Para poder acceder al software adecuado a nuestras necesidades tendremos que ser capaces de analizar cuáles son esas necesidades y cuál es nuestra disponibilidad de tiempo para aprender a usarlo satisfactoriamente. Hemos puesto cierto énfasis en la posibilidad de ser nuestros propios programadores de la herramienta que necesitamos.

Nuestra opinión es que no es una mala inversión para un ingeniero en disciplinas ajenas a la informática que dedique un tiempo a adquirir destrezas en programación, las que serán útiles en el campo del manejo de controladores en planta, análisis estadístico de datos y en otras aplicaciones.

Volviendo al recuento de alternativas que hemos hecho, vemos que los campos de aplicación de esas alternativas, (que son nuestras necesidades) se pueden dividir en tres grandes grupos:

- Resumen de datos e informes
- Estudios y ensayos
- Manejo de datos (Ciencia de los datos)

Resumen de datos e informes.

Para este tipo de tareas debemos disponer de representaciones gráficas, crear tablas de resumen de resultados, hacer cálculos pequeños, informes rápidos.

En estos casos, emplearemos hojas de cálculo y herramientas de reporte. Con ***Excel*** y/o ***Power BI*** nos debería alcanzar. Su nivel de uso es **<u>INTUITIVO</u>**.

Estudios y ensayos

Se aplica a las tareas de ingeniería en alimentos, biología, ingeniería industrial, estudios de mercado, etc. La característica es que se trabaja con relativamente pocos datos y se necesitan análisis estadísticos simples pero precisos. No hay necesidad de personalización de funciones.

Estos trabajos los podemos encontrar en la mayor parte de los artículos científicos. Es la esencia de lo que hemos tratado en este Libro 5 de ***Optimiza***.

Para satisfacer nuestras necesidades podremos usar, además de los del grupo anterior y para completar sus capacidades, programas dedicados o la posibilidad de programar algo simple. Así podemos pensar en ***MiniTAB***, ***SPSS*** o ***RCommander***. Su nivel de uso es de cierta **DIFICULTAD**.

Manejo de datos o Ciencia de datos.

Acá nos enfrentamos al manejo de grandes volúmenes de datos. Esto ocurre en Biología, en estudios de mercado, en análisis de indicadores económicos, en tecnologías de la información, en ingeniería de sistemas complejos como redes de comunicaciones o sistemas de control de máquinas complejas, etc.

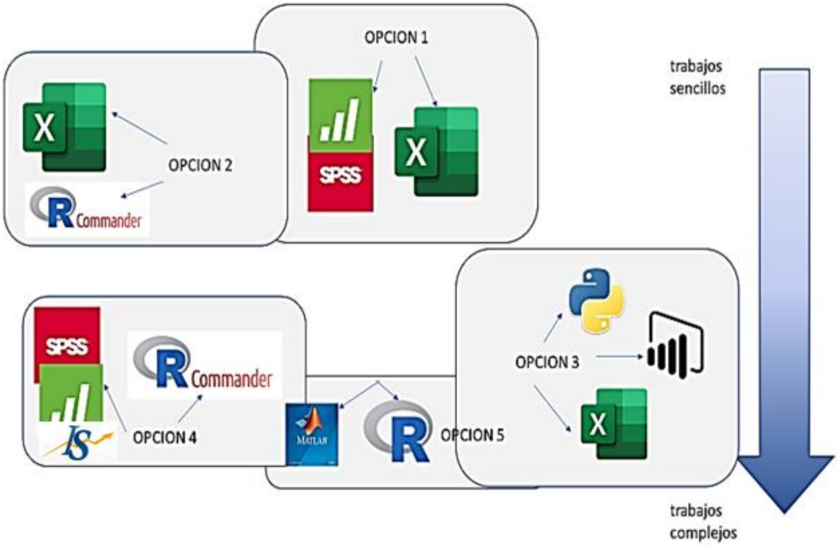
La lectura de datos debe ser segura y automatizada, ya que en general es imposible ingresar datos a mano. Los análisis deben ser personalizados. Los cálculos deben ser automatizados y soportar complejidad. Necesitaremos programar en ***MatLab***, ***R*** o ***Python***. Nivel de **GRAN DIFICULTAD**

Debemos notar que ***RCommander*** y ***RStudio*** usan como motor de cálculo el lenguaje ***R***, que se convierte así en la opción de software libre y que se puede acompañar con ***Excel*** para graficar, presentar informes y tablas.

En el siguiente gráfico mostramos las alternativas presentadas según la complejidad de uso y en función del tipo de destino que en general se les da a esos programas o de sus capacidades generales.

Insistimos en recordar que esta enumeración no es exhaustiva, solo mostramos algunas posibilidades que en nuestra opinión, son representativas y emergentes en un universo con muchas alternativas disponibles.

En este gráfico vemos distintas opciones de programas.



Referencias Bibliográficas del Libro 5

VIKTOR SCHÖNBERGER, CUKIER, K. "LA REVOLUCIÓN DE LOS DATOS MASIVOS". ED. TURNER. COLOMBIA, 2013

SALAS ARANDA, VICTOR "DISEÑO DE EXPERIMENTOS CON SOFTWARE ESTADÍSTICO". TF DE GRADO. UNIVERSIDAD DE SEVILLA, ESPAÑA, JUNIO 2018.

BRADLEY P.S., U.M. FAYYAD Y O. L. MANGARASARIAN, "MATHEMATICAL PROGRAMMING FOR DT MINING: FORMULATION AND CHALLENGES", JOURNAL OF COMPUTING, 11(3), 1999.

KUEHL, R. O. DISEÑO DE EXPERIMENTOS: PRINCIPIOS ESTADÍSTICOS DE DISEÑO Y ANÁLISIS DE INVESTIGACIÓN. THOMPSON LEARNING. MEXICO. MEXICO. 2001.

BOX, G. E. P., Y BEHNKEN, D. W. "SOME NEW THREE LEVEL DESIGNS FOR THE STUDY OF QUANTITATIVE VARIABLES." TECHNOMETRICS 2, 1966.

BOX, G. E. P., Y HUNTER, J. S. "MULTIFACTOR EXPERIMENTAL DESIGNS FOR EXPLORING RESPONSE SURFACES." ANNALS OF MATHEMATICAL STATISTICS 28. 1957

BOX, G. E. P., W. G. HUNTER Y J.S. HUNTER. "STATISTICS FOR EXPERIMENTERS. AN INTRODUCTION TO DESIGN, DATA ANALYSIS, AND MODEL BUILDING. WILEY. NEW YORK: 1978.

BOX, G. E. P., Y R.D. MEYER, R. D. "AN ANALYSIS FOR UNREPLICATED FRACTIONAL

BOX, G. E. P., Y K. G. WILSON, K. G. "ON THE EXPERIMENTAL ATTAINMENT OF OPTIMUM CONDITIONS." JOURNAL OF THE ROYAL STATISTICAL SOCIETY, B 13, 145. 1951

COCHRAN, W. G., AND COX, G. M. "EXPERIMENTAL DESIGNS", 2D ED. WILEY. NEW YORK. USA. 1957

Referencias Bibliográficas de la obra completa

HILLIER, F. S. Y G.J. LIEBERMAN. - INTRODUCCION A LA INVESTIGACIÓN DE OPERACIONES, 8 ED., MC GRAW HILL, MEXICO, MEXICO 2007

BRADLEY P.S., U.M. FAYYAD Y O. L. MANGARASARIAN, "MATHEMATICAL PROGRAMMING FOR DT MINING: FORMULATION AND CHALLENGES", JOURNAL OF COMPUTING, 11(3), 1999.

GASS, S. I., "MODEL WORLD: DANGER, BEWARE THE USER AS MODELER", INTERFACES, 20(3), 1990

PIDD, M., "JUST MODELING THROUGH: A ROUGH GUIDE TO MODELING", INTERFACES 29(2), 1999.

LINGO. "THE MODELING LANGUAGE AND OPTIMIZER". LINDO SYSTEMS PRESS. CHICAGO. USA, 2017

SCHARAGE, L. "OPTIMIZATION MODELING WITH LINGO". LINDO SYSTEMS PRESS. CHICAGO, USA. 2003

MAROS I. "COMPUTATIONAL TECHNIQUES OF THE SIMPLEX METHOD, KLUWER ACADEMIC PUBLISHERS, BOSTON, USA. 2003

HILLIER, F. S. Y M. S. HILLIER. "INTRODUCTION TO MANEGEMENT SCIENCE: A MODELING AND CASE STUDIES APPROACH WITH SPREADSHEETS" 2. ED. MC GRAW HILL, BURR RIDGE, IL. 2003

MURTHY, D. N. P., W. P. PAGE Y E. Y. RODIN, "MATHEMATICAL MODELING: A TOOL FOR PROBLEM SOLVING IN ENGINEERING, PHYSICAL, BIOLOGICAL AND SOCIAL SCIENCES. PERGAMON PRESS, OXFORD, INGLATERRA, 1990

WOLSEY, L. A., "STRONG FORMULATIONS FOR MIXED INTEGER PROGRAMS: VALID INEQUALITIES AND EXTENDED FORMULATIONS". MATHEMATICAL PROGRAMMING SERIES B, 97(1-2), 2003

ABBAS, A.E. Y R.A. HOWARD, “FOUNDATIONS OF DECISION ANALYSIS” AMAZON KINDLE ED. GLOBAL. PEARSON ED. ESSEX, INGLATERRA. 2016

KALASHNIVOV, V. V. “MATHEMATICAL METHODS IN QUEUING THEORY” SPRINGER SCIENCE & BUSINESS MEDIA, MOSCU, RUSIA. 2013

GUTIERREZ, V. Y VIDAL C.J. “MODELOS DE GESTIÓN DE INVENTARIOS EN CADENAS DE ABASTECIMIENTO: REVISIÓN DE LA LITERATURA”, REV. FAC. ING. UNIV. ANTIOQUIA N.° 43. ANTIOQUIA, COLOMBIA, 2008

KUEHL, R. O. DISEÑO DE EXPERIMENTOS: PRINCIPIOS ESTADÍSTICOS DE DISEÑO Y ANÁLISIS DE INVESTIGACIÓN. THOMPSON LEARNING. MEXICO. MEXICO. 2001.

BOX, G. E. P., Y BEHNKEN, D. W. “SOME NEW THREE LEVEL DESIGNS FOR THE STUDY OF QUANTITATIVE VARIABLES.” TECHNOMETRICS 2, 1966.

RIOS INSUA, D., S. RIOS INSUA, J. MARTIN JIMENEZ Y A MARTIN JIMENEZ. “SIMULACION METODOS Y APLICACIONES”, 2 ED. ALFAOMEGA RA-MA. MEXICO, MEXICO. 2009

BOX, G. E. P., Y DRAPER, N. R. “EMPIRICAL MODEL-BUILDING AND RESPONSE SURFACES”.: WILEY, NEW YORK. USA. 1987

BOX, G. E. P., Y HUNTER, J. S. “MULTIFACTOR EXPERIMENTAL DESIGNS FOR EXPLORING RESPONSE SURFACES.” ANNALS OF MATHEMATICAL STATISTICS 28. 1957

BOX, G. E. P., W. G. HUNTER Y J.S. HUNTER. “STATISTICS FOR EXPERIMENTERS. AN INTRODUCTION TO DESIGN, DATA ANALYSIS, AND MODEL BUILDING. WILEY. NEW YORK: 1978.

BOX, G. E. P., Y R.D. MEYER, R. D. “AN ANALYSIS FOR UNREPLICATED FRACTIONAL

BOX, G. E. P., Y K. G. WILSON, K. G. "ON THE EXPERIMENTAL ATTAINMENT OF OPTIMUM CONDITIONS." JOURNAL OF THE ROYAL STATISTICAL SOCIETY, B 13, 145. 1951

DANTZIG, G. B. Y M. N. THAPA, "LINEAR PROGRAMMING 1: INTRODUCTION" SPRINGER, NUEVA YORK, 1997.

COCHRAN, W. G., AND COX, G. M. "EXPERIMENTAL DESIGNS", 2D ED. WILEY. NEW YORK. USA. 1957

YOUNG P. Y S. ZAMIR. HANDBOOK OF GAME THEORY, VOLUME 4. (HANDBOOK). ELSEVIER. OXFORD, INGLATERRA, 2015.

LINDO. "USER'S MANUAL, LINDO SYSTEMS INC. PRESS, CHICAGO, USA. 2003

HOWARD, R.A., "THE ETHICAL OR/MS PROFESSIONAL", INTERFACES 31(6), 2001

LINGO USER'S GUIDE. LINDO SYSTEMS PRESS INC. CHICAGO, USA. 2003

FACTORIALS." TECHNOMETRICS 28. 1986

VIKTOR SCHÖNBERGER, CUKIER, K. "LA REVOLUCIÓN DE LOS DATOS MASIVOS". ED. TURNER. COLOMBIA, 2013

SALAS ARANDA, VICTOR "DISEÑO DE EXPERIMENTOS CON SOFTWARE ESTADÍSTICO". TF DE GRADO. UNIVERSIDAD DE SEVILLA, ESPAÑA, JUNIO 2018.