Parte I

Principios

Capítulo 1

Introducción

En este primer capítulo presentamos un breve esquema del estado del arte de nuestro tema de investigación. Asimismo, planteamos también la cuestión fundamental que intentaremos investigar: proponer una serie de criterios y consideraciones que puedan seguirse al estudiar superficies de respuesta que provengan de modelos no normales, concretamente, de aquellos cuya respuesta contenga datos binarios. Enfocamos el estudio de la probabilidad de éxito como respuesta en función de factores de variabilidad controlados por el experimentador. Para ello, elegimos el modelo logístico —ya sea como tal o bien utilizando la transformación logit— para diseños de hasta segundo orden, que definirán la clase de superficies de respuesta que estudiaremos.

1.1. Estadística y experimentación

A partir del estudio de lo que comúnmente se acepta como *sistema*[1] —especialmente, los relacionados con las ciencias naturales— en muchos casos ha sido posible llegar a formular una ley del tipo *universal* que describiera el funcionamiento del mismo, indicando mediante relaciones funcionales particulares de qué forma se relacionaban sus magnitudes. A través de la aplicación del método científico, este conocimiento fue validándose paso a paso en el transcurso del tiempo, hasta llegar a nuestros días sus versiones más actualizadas. Es así como a partir de ciertas hipótesis y expresiones formales —*fórmulas exactas* o *modelos matemáticos* y sus *condiciones de aplicabilidad*, o *condiciones de borde*— adecuadas al objeto de estudio, es el día de hoy que muchos fenómenos se pueden conocer de manera "razonablemente exacta", fruto de la validación de aquellas formulaciones de modo sistemático en los contextos para los cuales

[1] Utilizaremos este término cuando hagamos referencia a un determinado conjunto de *elementos relacionados entre sí*, como también a sus *interrelaciones*, los cuales se encuentran dentro de un *marco de estudio* determinado, y a las *fronteras* concretas y/o conceptuales que lo delimitan con el resto del todo lo que no forma parte del sistema, que podemos llamar *unvierso*.

se han definido aquellos modelos. A estos tipos de modelos suele dársele el nombre de *modelos mecanicistas*, en cuanto a que la relación que vincula las variables que representan el comportamiento del sistema bajo estudio, se dan de manera "mecánica" o "exacta"[2].

Sin embargo, no todos los sistemas pueden ser descritos con toda la precisión que el observador quisiera. Una explicación posible de esta afirmación se basa en un principio muy aceptado dentro del ámbito de la Estadística, según el cual, *no es posible inferir relación causal a partir de la consideración de modelos estadísticos solamente*[3]. Para que sea "razonablemente posible" hablar de relación causal entre un grupo de variables estadísticamente relacionadas, será necesario considerar otras formas adicionales de conocimiento, que permitan enriquecer el conocimiento proveniente de los modelos estadísticos considerados. Esto ocurre cuando el objeto de estudio está puesto sobre sistemas complejos, cuanto menos desde el punto de vista de la naturaleza que tienen los elementos que lo componen. Es la industria —por citar un ejemplo— un ámbito especialmente característico en donde los sistemas de modificación y/o de transformación de la materia no tienen patrones de comportamiento tan exactos como los que describen los modelos mecanicistas: por lo general es tan compleja la realidad de un proceso que resulta técnica y económicamente imposible definir con claridad "todas y cada una de las variables que componen el sistema", como así tampoco establecer "relaciones funcionales exactas entre sus variables".

Es por ello que la definición y estudio de este tipo de sistemas —ya sea en términos de sus elementos, variables o interrelaciones— será "inexacta" si intentamos compararla con la forma de definir los modelos mecanicistas. No obstante ello, muchas veces es posible estudiar estos sistemas por medio de restricciones a lo que sería el universo de posibilidades que existen alrededor de aquéllos. Una forma que se da frecuentemente en la práctica va de la mano de la experimentación con variables bajo el control del experimentador, de modo que al modificar los niveles de sus variables de manera criteriosa, es posible obtener información valiosa sobre el funcionamiento del sistema, representado en una o más variables que se pretenden controlar. En contraposición a los modelos mecanicistas, se está en presencia de otro tipo de modelos, de naturaleza característicamente estadística —y ya no "matemáticamente exacta"—, fruto de la experimentación

[2] Como ejemplo de sistemas mecanicistas, podemos citar a las famosas *Ecuaciones de Maxwell*, debidas al físico escocés James Clerk Maxwell (1831-1879), según las cuales, el comportamiento de sistemas electromagnéticos —sujetos a determinadas condiciones— puede ser descrito por un conjunto de cuatro ecuaciones integro-diferenciales. De esta manera, estas fórmulas forman parte de las llamadas *Leyes Universales del Electromagnetismo*, por ser de naturaleza universal en los fenómenos relacionados. La validez de las mismas, dentro de las condiciones en las cuales fueron definidas, puede considerarse "exacta".

[3] *Vid.* Box *et al.* (2005), p. 402; Kleinbaum *et al.* (1998), pp. 35 *et seq.,* Freedman *et al.* (1998), pp. 150 *et seq.*, entre otros.

racional con las variables que se piensa que tienen importancia dentro del sistema. Es muy frecuente encontrar que los modelos que describen esta clase de sistemas sean llamados *modelos empíricos*[4]: según ellos, será posible encontrar relaciones estadísticas útiles entre las variables que fueron consideradas al plantear el estudio, obtenidas éstas experimentalmente, y definidas para ciertas regiones de funcionamiento del sistema.

1.2. Construcción de modelos empíricos

El problema de construir modelos empíricos y de buscar en ellos sus valores extremos fue presentado formalmente a partir de los trabajos de Hotelling (1941) y de Friedman y Savage (1947), para ser formulado elegantemente en artículos como Box y Wilson (1951), en primer lugar, luego más desarrollados[5] en Box y Hunter (1957); Bradley (1958); Hunter (1954, 1956 y 1959) y Davies (1956), entre otros. A partir de estos desarrollos, que dieron en llamarse *Metodología de Superficie de Respuesta*, ha sido posible estudiar modelos empíricos que representen sistemas complejos desde un punto de vista estadístico integral y eficiente. En el trabajo de Box y Wilson (1951), los autores describen una metodología según la cual, a partir de la experimentación secuencial, es posible modelar determinado tipo de variables que se deseen controlar dentro de un sistema, a partir de la intervención del experimentador sobre sus niveles, de los que tiene control. En particular, el trabajo describe aquella metodología de diseño y de análisis cuando las variables a controlar tienen distribución normal de probabilidad.

Los detalles, aplicaciones y alcances de este enfoque fueron recibidos con éxito dentro del ámbito de la estadística experimental. El mismo fue recibiendo numerosas y valiosas aportaciones en los años sucesivos, tanto a nivel metodológico como de formas de realizar los cálculos. Al día de hoy, los resultados más autorizados y aceptados sobre la Metodología de Superficie de Respuesta se han publicado en referencias bibliográficas dedicadas al tema. Entre las más representativas, encontramos a: Myers (1976), Box y Draper (1987), Cornell (1990), Khuri y Cornell (1996) y Myers y Montgomery (2002), entre otros. En varios artículos muy interesantes también se describe en mayor grado el avance histórico-crítico del conocimiento en este campo específico, entre tantos otros, encontramos: Hill y Hunter (1966), Mead y Pike (1975), Myers *et al.* (1989) y Box (1999b). Entre éstos, uno de los más recientes en donde se hace una extensiva revisión bibliográfica muy completa es el de Myers *et al.* (2004).

Todas estas aportaciones han centrado su estudio suponiendo que la distribución

[4] *Vid.* p. ej. Box y Draper (1987).
[5] *Vid.* p. ej. Cornell (1990).

de la variable que se desea controlar es del tipo normal.

Cuando los supuestos de normalidad dejan de tener validez —característicamente, por el tipo de datos estudiados— hay ciertos puntos de aquella metodología que dejan de tener validez tal como fueron planteados. En lo que refiere al desarrollo de los diseños de experimentos para variables con distribución no normal, la formulación actual del problema parte de una serie de artículos y trabajos, entre los cuales podemos citar a: Abdelbasit y Plackett (1983); Bisgaard y Fuller (1994 y 1995); Myers *et al.* (1994); Hamada y Nelder (1997); Lewis (1998); Lewis *et al.* (1999, 2001a y 2001b); Bisgaard y Gerstbakh (2000) y Collett (2003), entre otros.

Sin dudarlo demasiado, el trabajo de Nelder y Wedderburn (1972) en donde presentan y formulan los *Modelos Lineales Generalizados*, ha resultado ser un valioso aporte metodológico para encontrar criterios sistemáticos y relativamente homogéneos a la hora de plantear modelos estadísticos de un número nada despreciable de distribuciones de probabilidad, cual es, la Familia Exponencial de distribuciones. En lo referente a los modelos no normales, este nuevo enfoque aporta una presentación muy sistemática y elegante que, desde el punto de vista del modelado estadístico, goza el día de hoy de un grado muy avanzado de madurez, y su aplicación se extiende cada vez más dentro del ámbito de las ciencias experimentales. Entre las obras más reconocidas que han sido dedicadas a estos modelos, debemos citar a: McCullagh y Nelder (1989); Lindsey (1997); McCulloch y Searle (2001); Fahrmeir y Tutz (2001); Myers *et al.* (2002); Olsson (2002); Hoffmann (2003) y Ato *et al.* (2005), entre otros, como así también un numero muy prolífico de artículos.

Parece natural pensar que el paso siguiente dentro de este proceso de avance del conocimiento fuese bastante directo: al igual que ocurre con el Modelo Lineal con la Metodología de Superficie de Respuesta, ésta también se encontrara desarrollada dentro del ámbito de los Modelos Lineales Generalizados. Sin embargo, hemos podido encontrar muy pocos antecedentes concretos que hablen de una "conciliación" entre la Metodología de Superficie de Respuesta y los Modelos Lineales Generalizados. Los trabajos que han intentado abordar el tema del Diseño y Análisis de Experimentos para modelos normales no lineales parten del trabajo seminal de Box y Lucas (1959), para encontrar muy pocas referencias en la actualidad que le hayan dado continuidad, como lo son Myers (1999); Lewis *et al.* (1999, 2001a y 2001b) y Myers *et al.* (2002). En particular, cuando el enfoque se extiende a la Metodología de Superficie de Respuesta —cuanto menos en el sentido completo en que la describen, por ejemplo, Box y Wilson (1951)— son muy escasos los trabajos que han abordado el tema. Entre ellos, hemos encontrado a Khuri (1993 y 2001) y Myers y Montgomery (2002), en los que se dan algunos puntos a considerar a la hora de poner el foco sobre los Modelos Lineales Generalizados en aras de intentar construir una Metodología de

Superficie de Respuesta.

Será entonces nuestra intención con la presente tesis, la de intentar estudiar nuevos enfoques que nos permitan conciliar la falta de enlaces concretos y formales entre estos dos grandes capítulos de la Estadística: extender los alcances de la Metodología de Superficie de Respuesta hacia los Modelos Lineales Generalizados. En particular, expondremos criterios de diseño para los modelos para Datos Binarios, que siguen distribuciones de Bernoulli o binomial.

1.3. Principios metodológicos

Para iniciar un intento por establecer criterios útiles que permitan extender los alcances de la *Metodología de Superficie de Respuesta* a los *Modelos Lineales Generalizados* —que abreviaremos de aquí en más como *MSR* y *MLG*, respectivamente— partiremos del enfoque clásico de la *MSR*, que considera variables con distribución normal. En el **Apéndice A** comentaremos algunos aspectos particulares de interés y utilidad metodológica para esta tesis referidos a la *MSR*.

Intentaremos resumir aquí la estructura desde la cual partiremos en aquel intento de conciliar ambos enfoques. Haciendo un paralelo con la Ingeniería de las Estructuras, diremos que con esta tesis utilizaremos las formas conocidas hasta el momento —es decir, los criterios y métodos de la *MSR*— como *guía de armado* de una nueva construcción, cuyos *elementos a ensamblar* serán los provenientes de los *MLG* y, en particular, de los *Modelos para Datos Binarios*[6], que abreviaremos de aquí en más como *MDB*. El criterio que seguiremos para preparar dichos elementos se apoyará sobre los principios del *Diseño de Experimentos*, los cuales tienen un carácter "estático", y encadenaremos unos elementos con otros desde la óptica de la *MSR*, adaptada a la naturaleza particular de sus elementos. Esto nos permitirá proponer un "enfoque dinámico" desde algunas estrategias que hemos elaborado, y que pretenderá ser nuestra aportación metodológica a estas áreas del conocimiento.

Como mencionáramos anteriormente, nos ha parecido razonable partir de una estructura existente para adentrarnos en aquella desconocida que pretendemos investigar, quizá por un motivo de respeto hacia quienes han sido los pioneros en estudiar sus fundamentos. Entonces, a modo de estructura de estudio, daremos un muy breve esquema de los elementos más característicos que componen la *MSR* para variables normales. Lo haremos sin entrar en mayores detalles, ya que las muy elocuentes referencias lo tratan de manera más que excelente: utilizaremos todo este conocimiento de

[6] Bajo este título, identificaremos en lo sucesivo a los datos del tipo *dicotómico*, ya sea para los que provienen de *Procesos de Bernoulli* como para las distribuciones *binomiales*. En el **Apéndice B**, resumimos los lineamientos fundamentales de las mismas, que utilizaremos a lo lago de todo este trabajo.

manera *estructural*.

Quedan claros los grandes propósitos tradicionales de la experimentación[7]:

- Explorar y cuantificar la relación funcional que existe entre los valores de una cierta variable susceptible de ser medida, o *respuesta*, y un cierto conjunto de variables controlables, o *factores*, de los que se presume que depende la respuesta[8], y
- Encontrar *qué niveles de estos factores* son los que producen el "mejor valor" de la respuesta de acuerdo con el objetivo del problema bajo estudio, en los que "mejor" puede significar *mayor*, *menor* o l*o más parecido posible* a un "valor target".
- Realizar *previsiones a futuro* acerca del valor se espera que tenga la respuesta para un nivel dado del conjunto de factores.

Aparece tradicionalmente la *MSR* como un conjunto de técnicas diseñadas para encontrar justamente aquel "mejor valor" de la respuesta de manera eficiente. De todos modos, como claramente se indica en Cornell (1990), si no fuera posible llegar a estos valores, mediante la experimentación será por lo general posible avanzar en el *conocimiento global del funcionamiento del sistema*, y en particular, de cómo es afectada la respuesta por los distintos factores considerados y los niveles que ellos tomen. Asimismo, podrá el experimentador ganar un conocimiento adicional en términos de la *región de operabilidad* y de la *región de experimentación* del sistema: la primera referirá a los valores adecuados[9] dentro de los posibles niveles que puedan tomar las variables, y la segunda, hará referencia a la región específica dentro de la que se realizarán las experimentaciones.

1.3.1. Fases secuenciales en la experimentación

Partiendo de reflexiones de Bates y Watts[10], coincidimos en que la experimentación resulta una metodología fundamental para el aprendizaje científico y técnico, que puede caracterizarse como un cierta *metodología sistemática de reducción de la ignorancia* sobre el funcionamiento de ciertos sistemas. En cada etapa de la investigación, el experimentador se encuentra con un cierto número de datos que utilizará

[7] Por el atractivo grado de síntesis y de claridad conceptual, tomaremos en esta sección algunas ideas que aparecen en Cornell (1990).

[8] Esto quedará de manifiesto a través del conocimiento que se consiga de los parámetros desconocidos del modelo, dados por un cierto vector $\boldsymbol{\beta}$.

[9] Mediante este término pretendemos indicar que el problema bajo estudio debería tener en cuenta —cuanto menos— una, varias o todas las siguientes factibilidades: a) técnica, b) económica, c) ética y/o moral, y d) medioambiental, entre otras.

[10] *Vid.* Bates y Watts (1988), p. 122.

para ir explicando gradualmente el sistema bajo estudio, de modo que a medida que avanza el experimento, se va consiguiendo una ganancia en el conocimiento de aquél o una reducción en la ignorancia de su funcionamiento.

Varios autores, entre ellos Box *et al.* (1978 y 2005) o Prat *et al.* (1997 y 2004) por citar algunos, suelen referirse a la llamada *estrategia secuencial de experimentación*, según la cual, se irá sucediendo uno tras otro una serie de experimentos controlados de modo que entre en cada paso sucesivo se vaya ganando la mayor cantidad de información posible sobre el sistema —generalmente desconocido— dentro del que se esté experimentando.

A este respecto, algunos autores[11] señalan tres grandes etapas o fases las cuales van sucediéndose a medida que se avanza en la experimentación. La prosecución de las mismas muestra una forma de avanzar de forma eficiente en la profundización del conocimiento del problema mediante la *MSR*. Estas fases son:

a. *Fase de "exploración experimental"*: permite averiguar qué factores deben considerarse como más importantes dentro de un conjunto mayor de factores "candidatos", de tal modo que su inclusión dentro del modelo sea significativa. En esta fase suelen emplearse los *diseños factoriales a dos niveles*, a partir de los cuales, un gran conjunto inicial de K factores quedará reducido a otro de tamaño k, $k < K$, con el cual se continuará experimentando posteriormente.

b. *Fase de "modelado empírico"*: en esta fase, una vez identificados los factores más relevantes que afectan a la respuesta, interesará saber cuán cerca o lejos se encuentran las condiciones actuales de experimentación del sistema de la región del óptimo (si lo hubiere). En esta fase, se utilizan modelos de primer orden y otros modelos de exploración "eficientes" de la superficie, como lo es el *Método de la máxima* (o *Mínima*) *pendiente de crecimiento*, o "steepest ascent (descent) method".

c. *Fase de "modelado mecanicista"*: en esta etapa final, ya en regiones experimentales cercanas al óptimo, se seguirá investigando qué forma tendrá la superficie de respuesta. mediante la consideración de modelos de segundo orden, se podrá tener un criterio adecuado de exploración de la curvatura de la superficie en zonas vecinas al óptimo (en caso de existir). Mediante el *análisis canónico*, será posible clasificar geométricamente aquella zona.

En este trabajo que preparamos, pretendemos hacer un "pool" de estas fases no demasiado riguroso en cuanto a los objetivos que persigue cada una, sino más bien de

[11] *Vid.* Box y Draper (1987), o bien Myers y Montgomery (2002), entre otros.

carácter flexible debido a que nos aventuramos con un problema relativamente nuevo para el que no sería provechoso ser demasiado ritualistas imponiendo objetivos rígidos.

En cuanto a los objetivos que persigue el experimentador durante este encadenamiento secuencial de experimentación, varios autores[12] coinciden en que el foco estará puesto, al menos básicamente, en uno o más de los siguientes puntos:

a. *Relación funcional*: define un modelo formal, expresado mediante una *fórmula explícita*, que muestra cómo se relacionan matemáticamente la respuesta y los factores más relevantes del caso estudiado, al menos, dentro de la región en la cual se ha experimentado.

b. *Especificación*: permite encontrar qué niveles de las variables explicativas satisfarán un determinado *valor-objetivo* (p. ej., un determinado nivel de producción), cumpliendo al mismo tiempo con conjunto de especificaciones.

c. *Exploración del punto estacionario*: consiste en encontrar qué niveles de los factores serán los que hagan que la respuesta alcance un valor óptimo local, llamado *punto estacionario*.

Es entonces que, mediante esta lógica *secuencial*, la definición progresiva de regiones de experimentación y la elección adecuada de los diseños a utilizar va permitiendo ampliar el conocimiento del sistema bajo estudio, en donde el experimentador tendrá que ir decidiendo:

- el grado de replicación de la respuesta que será necesario observar,
- la localización de las condiciones experimentales correspondientes al óptimo (si lo hubiere),
- la selección del tipo de función de aproximación más adecuada para cada etapa,
- la elección del tipo de diseño más adecuado, y
- la decisión de efectuar o no transformaciones en la respuesta.

1.3.2. Los diseños factoriales

Frente a la alternativa que muchas veces se plantea entre experimentar y analizar datos existentes, en ciertas ocasiones de desarrollo de nuevas ideas (ya sean procesos o productos nuevos), la primera alternativa debe tenerse muy en cuenta, ya que no

[12] *Vid.* p. ej. Box y Draper (1987); Khuri y Cornell (1996) y Myers y Montgomery (2002), entre otros.

siempre se encuentran disponibles antecedentes que puedan ser útiles a los fines de desarrollo. Varios autores sugieren que la segunda alternativa, es decir, el análisis de datos existentes, podría implicar que se incurra en algunos problemas, los cuales podrían, a su vez, conducir a posibles decisiones equivocadas. A este respecto, en Prat *et al.* (1997), pp. 127 *et seq.*, aparece una enumeración interesante sobre estos aspectos.

Si la opción elegida fuese la de experimentar, otro de los dilemas que suelen presentarse en estos casos es el de cómo realizar los experimentos. Por defecto, la primera opción sería la de realizar experimentos de forma intuitiva o sin planificar, mientras que una segunda opción sería la de pensar un poco más acerca de cómo realizarlos. Los mismos autores precitados hacen una breve discusión sobre las ventajas de los experimentos planificados versus los que no lo están[13]. A este respecto, nos hacemos eco, con los mismos autores, de una regla de oro de la experimentación: *"no invertir nunca todo el presupuesto en un primer conjunto de experimentos, y utilizar en su diseño toda la información disponible"*.

Los diseños factoriales tienen un uso muy difundido en los experimentos diseñados, particularmente en aquellos en donde intervienen varios factores de variabilidad y en los que es necesario examinar los efectos de los mismos sobre una respuesta[14]. La utilización de este tipo de diseños, tiene como idea básica la de realizar experimentos considerando la superposición de todos los efectos que tuvieran los factores considerados por el experimentador, como así también algunas funciones de dichos factores, como por ejemplo, la inclusión de términos cuadráticos, de interacciones de a dos factores, etc. Esta forma de realizarlos, los distingue de los llamados "diseños clásicos", en los que se suelen evaluar los efectos de a uno a la vez sobre la respuesta[15].

Ya un poco más avanzados y decididos a experimentar racionalmente que hacerlo de forma asistemática, vendrán las consideraciones del tipo estadístico para definir la estructura que tendrá el modelo que se utilizará. Esto es: de acuerdo a la naturaleza del problema a estudiar, se establecerán ciertas hipótesis acerca de cómo se piensa que se distribuyen los datos —entre otras: si tienen distribución normal o no, si los modelos lineales resultan adecuados o si será necesario definir los no lineales, etc.— y a partir de ello, definir la o las respuestas, los factores a considerar, los niveles que se le asignarán a éstos, y la forma con que se piensa que se relacionan las variables. Luego de la recogida de los datos y de las estimaciones de los efectos, el estudio suele concluir en la obtención de un modelo propuesto, para el cual se han validado las hipótesis iniciales, y el cual resultará útil para la predicción de nuevos valores bajo

[13] *Vid.* Prat *et al.* (1997), pp. 130 *et seq.*
[14] *Vid.* p. ej. Myers y Montgomery (2002).
[15] *Vid.* p. ej.: Peña (1998b), vol. 2, cap. 9, o Prat *et al.* (1997), pp. 131 *et seq.*

ciertas condiciones. Queda claro que puede darse el caso en que se obtenga más de un modelo propuesto.

Un aspecto particular en la experimentación lo constituye el *diseño*: esto es, decidir de antemano qué valores se les asignarán a los factores de tal forma que las conclusiones que se puedan extraer del modelo final validado, sean las mejores —o, al menos, las más confiables— desde el punto de vista de las implicaciones estadísticas que conlleva, por ejemplo, la predicción de nuevos valores para la respuesta. De forma básica, suele haber dos enfoques fundamentales: *a) asignar dos niveles a cada factor*, o *b) asignar más de dos niveles*. El número de experimentos a realizar para un número k de factores, cada uno de ellos con n niveles distintos[16], será igual a n^k, lo cual da una idea del esfuerzo experimental que conlleva la consideración de un número relativamente grande de factores y de niveles para éstos.

Por su sencillez y relativa facilidad en el uso, los diseños factoriales a dos niveles suelen ser los más utilizados, especialmente en la experimentación industrial. Entre otras, las siguientes razones fundamentan lo precitado[17]:

1. Proporcionan una excelente relación entre el esfuerzo experimental y la información obtenida;

2. Son sencillos de construir, realizar, analizar e interpretar; y

3. Son fáciles de combinar entre ellos para obtener diseños más complejos.

En los capítulos siguientes, abordaremos una síntesis de los diseños de primer orden sobre la base de los desarrollos existentes, en primer lugar para datos que sigan una distribución normal y en segundo, para los *MLG*. Como punto de partida útil, seguiremos la exposición lógica que se suele utilizar en los diseños factoriales para datos normales, cuya madurez ha quedado reflejada en la numerosa literatura existente al respecto. También ilustraremos brevemente las ideas mediante algunos casos típicos extraídos de la literatura en los que interesa estudiar una respuesta de naturaleza binaria (p. ej.: una proporción) en función de ciertos factores de variabilidad, los cuales se encuentran bajo el control del experimentador. En un principio, solamente ilustraremos los detalles más importantes del problema, aunque no los resolveremos sino hasta presentar en secciones siguientes un estudio más detallado sobre los diseños utilizados[18].

[16] *Vid.* p. ej. Prat *et al.* (1997 y 2004).

[17] *Ibídem.*

[18] Un aspecto importante que quisiéramos comentar es el que tiene que ver con la elección de los diseños factoriales como herramienta básica de nuestro estudio. Las buenas propiedades que poseen los diseños factoriales son bien conocidas, especialmente en el ámbito de los modelos lineales. Naturalmente, la exploración de otra clase de diseños también resultará muy valiosa para ampliar los alcances del tema.

1.3.3. La *MSR* y la experimentación secuencial

Una posible conceptuación útil sobre la estrategia que se sigue para que una experimentación resulte eficiente puede ser vista en un contexto de imágenes. El hecho de intentar construir un trozo de película a partir de imágenes individuales "criteriosamente escogidas y correlativas" resulta análogo al de intentar conocer el funcionamiento de un sistema —hecho que se manifiesta mediante el conocimiento de la relación funcional entre la respuesta y un conjunto de factores— a partir de un encadenamiento de diseños experimentales. Las imágenes que se tienen en consideración no serán otra cosa que los diseños utilizados durante la experimentación (p. ej.: diseños factoriales, diseños centrales compuestos, etc.). El paso que sigue a la elección de una imagen será decidido sobre la base de los resultados de pruebas estadísticas (p. ej.: pruebas de hipótesis), que definirá qué rumbo tomar en la definición de las siguientes condiciones de experimentación. Por supuesto, todo conocimiento complementario que se tenga sobre el sistema será necesario a la hora de la toma de decisiones de los pasos siguientes. Si la elección secuencial de las condiciones experimentales resulta exitosa, se podrá tener una buena aproximación de al menos una región —la región experimental— de la superficie de respuesta correspondiente, que será en cierto modo, representativa de la reconstrucción de un trozo de película a partir de imágenes individuales.

De esta forma, la naturaleza secuencial de la experimentación es una característica muy singular de la forma en que opera la lógica de la *MSR*: ir aumentando el conocimiento del sistema a medida que se va experimentando. Cuanto mejor se escojan las imágenes individuales —que serán los diseños— tanto más eficiente será la ganancia de conocimiento en los pasos sucesivos acerca del sistema. En este sentido, existen herramientas estadísticas muy útiles que permiten realizar esto de manera racional y eficiente.

1.4. Alcances de esta tesis

Es nuestra intención con este trabajo la de aportar un marco teórico y práctico novedoso, desde el cual extender los alcances de la *MSR* a los *MDB*, dejando trazada alguna línea de investigación que resulte útil para explorar otros modelos de la familia exponencial de distribuciones. En este ambicioso objetivo, partimos de las aportaciones previas más importantes que se han realizado hasta el momento, y tratamos de conducir el estudio hacia criterios que sean aplicables en ciertos problemas de la realidad, en los que se manejen datos provenientes de distribuciones de Bernoulli y binomial. A este propósito, discutiremos la definición de un par de estrategias que hemos definido para valorar los resultados obtenidos: una de ellas está basada en la *cantidad de in-*

formación del modelo ajustado, mientras que la otra tiene en cuenta un *criterio de aproximación al máximo* del proceso bajo estudio.

1.4.1. Objetivo primario

Enlazando los dos grandes puntos precitados, es decir, la *MSR* como estrategia secuencial y los *MLG* como herramienta para estudiar modelos, el foco de este trabajo se centrará alrededor de definir una metodología que pretende estudiar cómo se pueden aprovechar los conocimientos actuales sobre *MSR* para datos con distribución normal y explorar su extensión a los datos binarios, vistos éstos como un caso especial de los *MLG*.

Puesto que tanto el *diseño* como el *análisis* de experimentos tienen objetivos diferentes y complementarios, pondremos nuestro énfasis en el primero de los aspectos, el diseño, de modo de estudiar con detenimiento su pregunta clásica: *dónde y cómo seleccionar los puntos del espacio de los factores de modo más adecuado para obtener la máxima información del proceso y conseguir alcanzar el objetivo del problema de la forma más eficiente posible.* Nos apoyaremos en este sentido en el enfoque secuencial de experimentación mediante sucesiones de diseños factoriales. En cuanto al segundo aspecto, el análisis, y dado que se encuentra en un estado de evolución mucho más avanzado que el de diseño[19], solamente abordaremos algunos puntos seleccionados, cuya profundidad y desarrollo exceden el alcance que nos hemos propuesto en este trabajo.

1.4.2. Consideraciones específicas de estudio

Las funciones de aproximación de la respuesta

Si denotamos mediante $f(\boldsymbol{\xi}, \boldsymbol{\theta})$ a la relación funcional —en un principio, desconocida— que suponemos que existe entre la respuesta verdadera, $\mathbf{y}$, y un conjunto de factores asociados, $\boldsymbol{\xi}$, en nuestro contexto de estudio nos interesará encontrar una cierta función de aproximación $g(\mathbf{x}, \boldsymbol{\beta})$, de tal modo que la misma resulte una buena aproximación local de $f(\boldsymbol{\xi}, \boldsymbol{\theta})$ en una cierta región experimental, es decir, para que se cumpla que $f(\boldsymbol{\xi}, \boldsymbol{\theta}) \simeq g(\mathbf{x}, \boldsymbol{\beta})$, en donde $\boldsymbol{\theta}$ y $\boldsymbol{\beta}$ son vectores de parámetros desconocidos, asociados a los factores de variabilidad.

Cuando existen evidencias según las cuales la respuesta puede considerarse que siga una distribución normal, el problema ya se encuentra resuelto, mientras si no sigue esta distribución, aún no hemos encontrado una solución integral tan completa como en el primer caso. En particular, aquí nos centraremos en explorar qué consecuencias

[19] *Vid.* Lewis *et al.* (1999, 2001a y 2001b); Hamada y Nelder (1997) y Myers *et al.* (2002), entre otros.

se siguen cuando los datos son de naturaleza binaria. En los **Apéndices A** y **B** de este trabajo presentamos algunos elementos de interés al respecto, que fundamentan muchos de los conceptos que desarrollamos en esta sección.

Respuestas binarias y modelos para la probabilidad de éxito

Mientras que en el modelo lineal normal se buscan funciones polinómicas para aproximar linealmente la respuesta en función de los factores, intentaremos explorar en este trabajo otras funciones de aproximación para la *probabilidad de éxito*[20]. Consideraremos entonces dicha probabilidad como función de un conjunto de factores dado por el vector $\mathbf{x} = (x_1, x_2, \ldots, x_k)'$, a partir del cual se define un polinomio llamado *predictor lineal* del modelo, que depende también de un conjunto de parámetros desconocidos, $\boldsymbol{\beta}$, y de manera lineal. Para el caso particular de considerar $k = 2$ factores, este predictor lineal será de la forma:

$$\eta = \beta_0 + \sum_{j=1}^{2} \beta_j x_j + \sum_{i<j}^{2} \beta_{ij} x_i x_j + \sum_{j=1}^{2} \beta_{jj} x_j^2 \tag{1.1}$$

o bien de modo vectorial:

$$\eta = \beta_0 + \mathbf{x}'\mathbf{b} + \mathbf{x}'\mathbf{B}\mathbf{x}, \tag{1.2}$$

en donde:

$$\mathbf{b} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \quad \text{y} \quad \mathbf{B} = \begin{bmatrix} \beta_{11} & \frac{1}{2}\beta_{12} \\ \frac{1}{2}\beta_{12} & \beta_{22} \end{bmatrix}. \tag{1.3}$$

Resumiendo lo antedicho, tomaremos un modelo conveniente para aproximar la función probabilidad de éxito, que dependerá del conjunto de factores elegidos, $\mathbf{x}$, y del vector de parámetros $\boldsymbol{\beta}$, y que denotaremos como $\pi(\mathbf{x}, \boldsymbol{\beta})$. Definiendo la variable aleatoria de Bernoulli, que denotaremos por y, y asignando un valor de $y = 1$ al resultado "éxito", aproximaremos la misma mediante la siguiente expresión:

$$P(y = 1 \mid \boldsymbol{\xi}, \boldsymbol{\theta}) \simeq \pi(\mathbf{x}, \boldsymbol{\beta})$$

De acuerdo con esto, la función de aproximación de la probabilidad de éxito será considerada como una relación funcional que dependerá de un conjunto de factores $\mathbf{x}$ y de otro conjunto de parámetros $\boldsymbol{\beta}$. Entre varios posibles modelos, hemos elegido el llamado *modelo logístico*[21] para definir esta relación convenientemente, de modo que la misma contenga tanto los factores como los parámetros, los cuales habrá que

[20] Sobre este tipo de distribuciones, comentaremos algunos aspectos complementarios en el Apéndice C.

[21] Más adelante se verá que es posible partir de otros modelos distintos al logístico para aproximar la probabilidad de éxito, como por ejemplo, el modelo *probit*, el *"complementary log-log"*, entre otros.

estimar en su momento. Dicha relación define el *modelo logístico para la probabilidad de éxito*[22], que será una expresión del tipo:

$$\pi\left(\mathbf{x},\boldsymbol{\beta}\right)=\frac{\exp(\beta_0+\mathbf{x}'\mathbf{b}+\mathbf{x}'\mathbf{B}\mathbf{x})}{1+\exp(\beta_0+\mathbf{x}'\mathbf{b}+\mathbf{x}'\mathbf{B}\mathbf{x})}, \tag{1.4}$$

en la que la cantidad $\eta=\beta_0+\mathbf{x}'\mathbf{b}+\mathbf{x}'\mathbf{B}\mathbf{x}$ representa una característica importante[23] de este modelo, cuyos detalles comentaremos en capítulos siguientes. En el **Apéndice B** comentaremos también otros detalles sobre este modelo.

Mediante un sencillo despeje algebraico de la ecuación (1.4), queda definida la *transformación logit* del mismo modelo logístico, llegándose a la siguiente expresión equivalente:

$$\text{logit}\ \left[\pi\left(\mathbf{x},\boldsymbol{\beta}\right)\right]=\ln\left[\frac{\pi\left(\mathbf{x},\boldsymbol{\beta}\right)}{1-\pi\left(\mathbf{x},\boldsymbol{\beta}\right)}\right]=\beta_0+\mathbf{x}'\mathbf{b}+\mathbf{x}'\mathbf{B}\mathbf{x} \tag{1.5}$$

Siguiendo con lo que veníamos indicando sobre las funciones de aproximación de la respuesta verdadera —en nuestro caso, la misma es $\pi\left(\mathbf{x},\boldsymbol{\beta}\right)$— tomaremos la misma transformación logit como función de aproximación, es decir:

$$g\left[\pi\left(\mathbf{x},\boldsymbol{\beta}\right)\right]=\text{logit}\ \left[\pi\left(\mathbf{x},\boldsymbol{\beta}\right)\right]$$

Queda claro que el modelo logit no es un modelo nuevo ni distinto al logístico, sino que es una forma conveniente de expresar el segundo. La nota característica que puede apreciarse al aplicar esta transformación es que se obtiene una función lineal en los parámetros —el predictor lineal— , para la cual será muy provechosa la teoría existente acerca del modelo lineal, cuyo rango de validez ya no tendrá las cotas fijas que tiene el modelo logístico[24].

Estrategias de experimentación

De acuerdo con lo mencionado anteriormente, nos moveremos dentro de la lógica secuencial de experimentación para conocer las características generales de las superficies ajustadas que se vayan obteniendo en los pasos sucesivos.

[22] *Vid.*: Collett (2003), Hosmer y Lemeshow (2000) o Kleinbaum (1994), entre otros.

[23] Como veremos en el **Apéndice B**, este constituye el *predictor lineal* o *componente sistemática del modelo*, característica muy distintiva de los *MLG*. Mientras que el predictor lineal puede tomar cualquier valor dentro del intervalo real $(-\infty,\infty)$, la función $\pi\left(\mathbf{x},\boldsymbol{\beta}\right)$ permanecerá acotada entre 0 y 1, lo cual permite asociarle una probabilidad. (No lo demostraremos aquí, pero queda claro se cumplen también las condiciones tanto axiomáticas de Kolmogorov como las de espacio probabilizable relacionados con las $\sigma-$álgebras de Borel y otras condiciones para que tenga sentido hablar formalmente de probabilidad).

[24] Profundizando un poco más lo anterior, y enlazándolo con los modelos no lineales, podemos terminar este párrafo agregando que $\pi(\mathbf{x},\boldsymbol{\beta})$ no es función lineal de $\mathbf{x}$ ni de $\boldsymbol{\beta}$, aunque ello no obstante, su "no linealidad" es conocida. La transformación logit es la manera de linealizar una no linealidad conocida en este modelo, en contraposición a la construcción de modelos empíricos del tipo no lineal, más en el estilo de modelado que sigue la línea de Bates y Watts (1998), por ejemplo, en que por lo general no siempre es posible linealizar los modelos.

A medida que vayan generándose nuevos diseños, aprovecharemos toda la información contenida en éste y en todos los diseños anteriores para ajustar los sucesivos modelos, en los cuales emplearemos tantos términos en el predictor lineal como nos lo permita el número de puntos de soporte disponibles. Por ejemplo, si en un primer diseño disponemos de $n = 5$ puntos experimentales, ensayaremos ajustes de modelos con un máximo de $p = 5$ parámetros. En nuestro caso, manejaremos modelos completos de segundo orden, con $p = 6$ parámetros. Si en un segundo diseño tomamos otros 5 puntos más, entonces el segundo modelo propuesto consistirá en $p = 6$ parámetros que ajusten un conjunto de $n = 10$ puntos experimentales, disponiendo así de $n - p = 4$ grados de libertad para determinar una medida muestral de la variabilidad de los datos, que estará representada por la función *devianza*, de la que profundizaremos más adelante.

En todos los casos, se tratará en la medida de lo posible de aprovechar la flexibilidad de paquetes informáticos existentes para desarrollar aplicaciones que permitan sistematizar los cálculos y comprobaciones que vayan sucediendo. En nuestro caso, y debido a las amplias posibilidades que ofrece, hemos elegido **R** (v. 2.2.1) como entorno de programación y cálculos estadísticos, del que describiremos en el **Apéndice C** las principales líneas de programa que utilizamos para nuestros cálculos.

1.4.3. Algunas de nuestras preguntas de investigación

Algunas de nuestras inquietudes sobre los alcances actuales de la *MSR* y de los *MLG* que nos han impulsado a realizar este trabajo los podemos resumir mediante las siguientes preguntas, entre otras:

a. ¿Cómo deberíamos simular un proceso de generación de datos que siga una distribución binomial?

b. ¿Qué métodos deberíamos seguir para ajustar modelos a los puntos simulados?

c. ¿Qué criterios pueden seguirse para seleccionar los términos de los modelos ajustados?

d. ¿Qué punto puede tomarse por válido como primer centro de experimentación?

e. ¿Qué niveles pueden considerarse para los factores estudiados?

f. ¿Qué longitud debe considerarse como "salto" entre diseños sucesivos?

g. ¿Cómo se determinan los "steepest paths" en los sucesivos modelos ajustados?

h. ¿Qué medidas de calidad de ajuste pueden definirse para evaluar los diseños estudiados?

i. ¿Es posible definir herramientas informáticas que permitan estudiar y facilitar los cálculos sistemáticamente?

A lo largo de las páginas que siguen nos hemos propuesto mostrar lo que hemos podido aprender como resultado de dedicar un gran esfuerzo en estudiar estos procesos y sus metodologías de mejora, y hemos tratado de aproximar lo mejor posible algunas respuestas a las preguntas precitadas. Tal vez el mayor mérito de todo esto no sea tanto el avance del conocimiento exclusivamente, sino el haber podido conseguir algunas aportaciones fruto de trabajar en equipo y en buena sintonía conceptual, profesional y humana. Creemos que el profesional en general —y en particular, el ingeniero— puede ser un gran habilidoso en el planteo de problemas muy complejos, en encontrar modelos útiles[25] que los representen y en implementar soluciones muy eficientes. Sin embargo, creemos más aún que su principal misión es la de *contribuir a la conducción de personas* hacia objetivos adecuados y realizables, de la forma que mejor convenga, teniendo siempre como centro a la persona humana como verdadero artífice del cambio y no a las fórmulas, los modelos y sus controles, que si bien pueden ser buenas herramientas de trabajo, por sí solas nunca alcanzarán su mejor integración en cualquier sistema que involucre tanto recursos materiales como personas.

[25] Para cerrar este capítulo, nos resulta inevitable escapar a la cita de la célebre reflexión atribuida al Prof. Greorge E. P. Box: *"all models are wrong, but some of them are useful"*.