TEMA 1: Diseño y análisis de experimentos

En prácticamente todos los campos de la ciencia y la ingeniería se realizan de manera habitual experimentos orientados a adquirir conocimiento sobre un proceso o sistema relevante. De una forma u otra, un experimento consiste en una prueba o test en el que se realiza la medición de alguna característica de interés. Más formalmente, es posible definir un experimento como una prueba o serie de pruebas en que se realizan **cambios intencionados en un conjunto de factores** o variables de entrada del sistema de cara a observar e identificar las **causas de la variabilidad** en una **variable respuesta** de interés.

Un sistema puede ser representado como se muestra en la Figura 1.1, esto es, como un mecanismo complejo que transforma un input (por ejemplo, materias primas) en un output (por ejemplo, un producto) con una o más variables respuesta observables. Este mecanismo se desarrolla en unas condiciones determinadas por una serie de factores, algunos de los cuales pueden ser controlables por el experimentador, mientras otros serán típicamente no controlables.

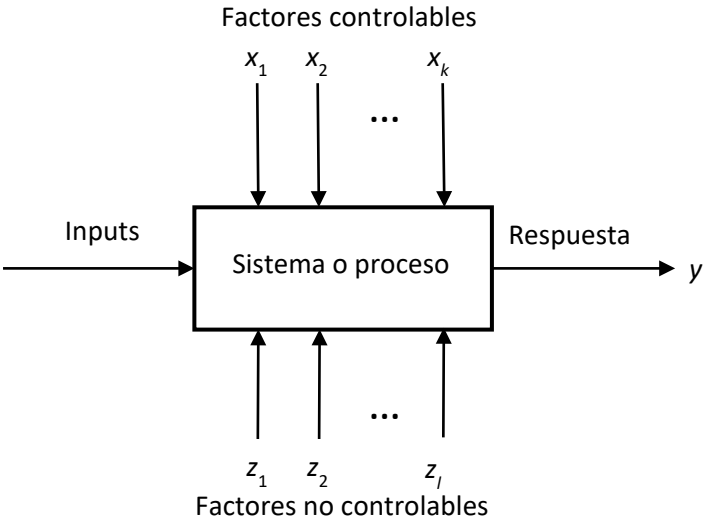

Figura 1.1: Representación de un sistema o proceso

Los objetivos de un experimento realizado sobre un sistema o proceso pueden incluir los siguientes:

1. Determinar qué factores tienen mayor influencia en la respuesta *y*.
2. Determinar qué valores han de tomar los factores influyentes de modo que la respuesta *y* tome un valor objetivo.
3. Determinar bajo qué valores de los factores influyentes se obtiene una variabilidad reducida de la respuesta *y*.
4. Determinar bajo qué valores de los factores controlables influyentes se minimiza la influencia de los factores incontrolables.

Para llevar a cabo estos experimentos de una manera eficiente y que permita el cumplimiento de estos objetivos, debe emplearse un enfoque científico en la planificación y realización del experimento y en el análisis de los datos resultantes de este. El **diseño y análisis estadístico de experimentos** es la materia que proporciona este enfoque científico, y se refiere al proceso de planificar el experimento de manera que se recojan datos adecuados susceptibles de ser analizados estadísticamente y de modo que sea posible extraer conclusiones válidas y objetivas.

1.1. CONCEPTOS BÁSICOS

Para ayudar a introducir las nociones e intuiciones básicas, supóngase el siguiente experimento ilustrativo: en un horno metalúrgico se está interesado en comparar el efecto de dos o más procesos de forjado en la dureza de una nueva aleación. El objetivo es determinar qué proceso proporciona la mayor dureza de esta aleación. Para ello se decide producir un número de especímenes (por ejemplo, unas pequeñas placas) de la aleación mediante cada proceso en estudio, y luego medir su dureza. En base a estas mediciones, es entonces posible obtener una dureza media de los especímenes producidos mediante cada proceso, y la idea es utilizar estas medias de cara a decidir qué proceso es mejor.

Así pues, a primera vista podemos entender este ejemplo como un sistema en el que: i) se seleccionan unas determinadas **unidades experimentales**, en este caso ciertas cantidades de materias primas para realizar los especímenes de la aleación; ii) se somete a diferentes conjuntos de esas unidades a diferentes **condiciones experimentales**, en este caso la producción de especímenes de la aleación mediante los diferentes procesos de forjado en estudio; y iii) se realiza la medición de una característica o **respuesta** de interés en las diferentes unidades experimentales, en el ejemplo una medición de la dureza de cada espécimen. La idea es observar cómo varía esta respuesta medida (la dureza de los especímenes) al variar esas condiciones (el tipo de proceso que los produce). Nótese que en este caso las condiciones vienen en principio determinadas por un único **factor**, el tipo de proceso empleado, que es controlable (el experimentador puede decidir qué proceso aplicar a cada paquete de materias primas con el que producir un espécimen), y diríamos que este factor tiene tantos **niveles** o valores posibles como procesos de producción se desea comparar.

En una situación ideal, toda la **variabilidad observada en la respuesta** sería debida a la variación en los niveles del factor de interés, de modo que podríamos concluir sin más que el proceso que proporciona una mayor dureza media sería el mejor. De hecho, en una situación perfecta no necesitaríamos producir más de un espécimen con cada proceso, ni siquiera calcular medias, ya que cada proceso produciría siempre especímenes con la misma dureza. Sin embargo, la realidad suele estar bastante alejada de esta situación ideal.

En este sentido, es importante entender que en cada proceso o sistema real normalmente operan muy **diversas fuentes de variabilidad**. En el caso que nos ocupa, una primera causa de variabilidad en las durezas medidas puede ser efectivamente la variación en los niveles del factor *X* = *tipo de proceso de forjado*. Esta variabilidad es deseada, en tanto es justo lo que se desea estudiar y es producida por los cambios controlados que realiza el experimentador en las condiciones en que opera el sistema. Denominaremos a este tipo de variabilidad como **variabilidad debida a las condiciones de interés**, o también variabilidad planificada.

Una segunda fuente de variabilidad estará típicamente asociada a pequeñas **variaciones aleatorias y no controlables en la realización del experimento**, y en particular del procedimiento mediante el que se realizan las mediciones de la respuesta, que en la práctica casi nunca estará libre de producir **errores de medición**, normalmente con un comportamiento aleatorio. Dos especímenes idénticos producidos mediante el mismo proceso con frecuencia (o incluso casi siempre) obtendrán valores medidos de dureza no exactamente iguales, debido a diferencias en el uso del sistema de medición, la precisión o calibrado de este, su desgaste, la realización de la medición en condiciones no 100% idénticas, etc. Incluso cuando el experimento utilice un procedimiento de medición de la respuesta totalmente fiable, dos especímenes producidos mediante el mismo proceso no tendrán exactamente la misma dureza, ya que el proceso no será nunca exactamente el mismo: pequeñas variaciones en la temperatura del horno, el tiempo empleado, las concentraciones usadas en la aleación o la calidad de las

materias primas introducirán en las mediciones de dureza cierta variabilidad no debida al factor que se está controlando. Este tipo de variabilidad, que llamaremos **variabilidad aleatoria, error experimental o ruido**, es no deseada y difícilmente evitable, aunque como veremos después determinados procedimientos experimentales pueden aliviar su efecto sobre las conclusiones del experimento.

Una tercera causa de variabilidad en la respuesta observada suele aparecer a causa de la **influencia sistemática, no aleatoria, de factores no controlables o no planificados**. El gran problema con este tipo de variabilidad es que pase inadvertida para el experimentador previamente a la planificación y realización del experimento. Por ejemplo, supóngase que se están comparando dos procesos de forjado, A y B, que precisan de cierta experiencia y práctica para llevarlos a cabo, y el proceso A lo realiza siempre el mismo operador, digamos O, encargado de producir todos los especímenes requeridos con ese proceso, e igualmente el proceso B lo realiza siempre un mismo operador P, diferente a O. Entonces, si se observa que los especímenes realizados mediante el proceso A proporcionan una dureza mayor que los del proceso B, o *vice versa*, no es posible discernir a ciencia cierta si esta mayor dureza es debida a una mayor eficacia del proceso A respecto al B o a una mayor habilidad del operador O en comparación con P. Lo mismo sucede si las materias primas utilizadas para fabricar los especímenes del proceso A provienen de un lote L, y las usadas en el proceso B provienen de un lote diferente M: en este caso, quizás las diferencias de dureza observadas sean debidas a diferencias de calidad en los lotes más que a la diferencia entre los procesos. Y algo similar puede ocurrir si, al producir en un mismo horno los n especímenes requeridos para cada proceso, se llevan a cabo primero todos los del proceso A y justo a continuación todos los del B. Si las condiciones del horno sufren pequeños cambios a medida que se van produciendo especímenes, las condiciones medias en que se producen los del proceso A, los n primeros, pueden llegar a ser muy diferentes a las condiciones medias en que se producen los del B, los n últimos, pudiendo llevar a observar diferencias de dureza importantes entre procesos de eficiencia similar, o que no existen diferencias entre procesos de distinta eficiencia. En general, este tipo de **variabilidad sistemática no planificada** puede arruinar todo el trabajo y recursos empleados en la realización de un experimento, en tanto introduce **sesgo** en la estimación del efecto real del factor de interés, haciendo variar sistemáticamente hacia un lado u otro un conjunto de mediciones sin que seamos capaces de reconocer la causa o incluso que este sesgo está presente.

Afortunadamente, existen **procedimientos experimentales** y de análisis estadístico que permiten aliviar el efecto de la variabilidad aleatoria, así como convertir la variabilidad potencialmente no planificada en variabilidad aleatoria, o incluso en variabilidad planificada, siempre que se lleve a cabo un correcto diseño del experimento.

Uno de estos procedimientos, de importancia capital a pesar de su aparente simplicidad, consiste en la **aleatorización** del experimento, que se define formalmente como la asignación aleatoria de unidades experimentales a los diferentes niveles del factor y realizaciones del experimento. En general, la aleatorización permite proteger el experimento contra la introducción de variabilidad sistemática producida por factores no controlables, ya que las desviaciones que estos producen tenderán en promedio a ser las mismas en cada nivel del factor cuando el orden de realización de las pruebas sea aleatorio. Así, si en el ejemplo anterior el orden en que se producen los $2n$ especímenes totales mediante ambos procesos fuese aleatorio, de modo que para cada paquete de materias primas con que realizar un espécimen (la unidad experimental) se sortease el proceso que lo llevará a cabo, entonces el efecto sobre la dureza debido a las condiciones cambiantes del horno afectaría de manera muy similar a ambos procesos de forjado, eliminando el sesgo hacia uno de los procesos que antes se describía. En

este sentido, la aleatorización transforma la variabilidad sistemática en variabilidad aleatoria, que presenta un menor riesgo para la validez de las conclusiones del experimento.

Un segundo mecanismo experimental para evitar la aparición de variabilidad sistemática es la utilización de **variables bloque**, que permiten que la variabilidad debida a factores controlables que no son de interés pueda ser tratada como variabilidad planificada, reduciendo así el error experimental y eliminando el posible sesgo causado por tales factores. En el ejemplo anterior, los diferentes operadores que llevan a cabo el proceso de forjado podrían considerarse como niveles de una variable bloque denominada *operador*. La utilización de bloques requiere que los diferentes niveles del factor de interés estén **cruzados** con los niveles de la variable bloque, esto es, que en cada nivel del bloque se asignen aleatoriamente unidades experimentales a cada nivel del factor de interés. En el ejemplo, esto implicaría que cada operador, O y P, han de realizar especímenes mediante cada uno de los dos procesos A y B. De este modo, si el operador O es efectivamente más hábil que el operador P, este procedimiento permite no solo repartir este sesgo entre ambos procesos, sino también detectar esta diferencia y sustraer su efecto de la variabilidad observada en la dureza de los especímenes.

Un tercer y último mecanismo para reducir el error experimental consiste en la **replicación o repetición del experimento**. Al aumentar el número n de mediciones efectuadas, es posible calcular promedios más precisos. Esto es debido a que la varianza de la media de un conjunto de observaciones es inversamente proporcional al número de observaciones, pues si la varianza de cada observación es σ^2, entonces la varianza de la media es σ^2/n. No obstante, es importante entender que la replicación del experimento conlleva necesariamente un aumento del coste económico asociado, en tanto es necesario emplear más material experimental, energía, trabajadores, etc.

Una pregunta que puede surgir después de introducir estos conceptos básicos sobre fuentes de variabilidad es la siguiente: ¿Por qué tienen tanta importancia estas diferentes fuentes de variabilidad de la respuesta en la realización y análisis de un experimento y en las conclusiones que se pueden extraer de él?

Una primera clave es que, como se ha ilustrado al describir la variabilidad sistemática no planificada, existe una fuerte **dependencia entre los resultados y conclusiones que se pueden extraer de un experimento y el modo en que este ha sido realizado** y los datos que se han recogido. Determinados modos de llevar a cabo el experimento no permitirán separar analíticamente algunas de estas fuentes de variabilidad, lo que conducirá habitualmente a conclusiones erróneas o a dificultades para demostrar el efecto de interés bajo estudio. Por ello, es fundamental una correcta identificación de estas fuentes de variabilidad y diseñar el experimento de manera adecuada para afrontarlas.

Para ilustrar esta dependencia entre el diseño de un experimento y las conclusiones que es posible obtener de él, supongamos que en el ejemplo anterior se produce un único espécimen de aleación mediante cada proceso de forjado, de manera que tras realizar las mediciones se observa que la dureza del espécimen del proceso A es 6.0 y la del espécimen del proceso B es 8.0. ¿Podemos estar seguros en estas condiciones de que el proceso B es mejor que el A? Claramente no, pues estos datos no nos permiten **estimar la variabilidad aleatoria inherente al proceso**. No tenemos ninguna garantía de que, al producir otro par de especímenes en condiciones idénticas, uno con cada proceso, los nuevos resultados no sean de algún modo opuestos a los anteriores: por ejemplo, que ahora la dureza observada del espécimen A sea 8.0 y la del B sea 7.0. En otras palabras, las diferencias de dureza observadas podrían ser debidas al error experimental de ambos procesos más que a una diferencia real de eficacia entre ellos, y

no es posible detectar o descartar esto si solo se realiza una medición en cada proceso. En estas condiciones, es imposible extraer conclusiones útiles del experimento.

Este ejemplo pone de manifiesto una segunda clave en relación a la importancia de conocer y aislar las diferentes fuentes de variabilidad. De cara a poder concluir que el factor en estudio (el tipo de proceso de forjado) tiene un efecto real sobre la respuesta (la dureza de la aleación), es necesario **comparar la variabilidad atribuible al efecto del factor bajo estudio con la debida al error experimental**. Siguiendo con el ejemplo anterior, tras la segunda observación de cada proceso, se tienen las mediciones de dureza 6.0 y 8.0 para el proceso A y 8.0 y 7.0 para el proceso B. Así pues, la media observada para el proceso A es 7.0 = (6.0 + 8.0) / 2, y para el proceso B es 7.5 = (8.0 + 7.0) / 2. Esto arroja una diferencia de 0.5 entre el efecto atribuible al nivel A del factor y el atribuible al nivel B. Sin embargo, esta diferencia de 0.5 entre los efectos estimados de los niveles del factor parece relativamente pequeña en comparación con la variabilidad que se observa en las mediciones obtenidas dentro de cada nivel, que sin mayores detalles sobre el diseño del experimento solo es atribuible al error experimental. En otras palabras, aunque la media observada de un proceso es mayor que la del otro, la magnitud observada del error experimental no permite en este caso asegurar que esas diferencias entre las medias de ambos niveles sean debidas a una diferencia real de eficacia entre ambos procesos más que a las variaciones inherentes a ellos. Sin embargo, si las mediciones de dureza realizadas hubieran sido de 6.9 y 7.1 para el proceso A y 7.4 y 7.6 para para el proceso B, se tendrían de nuevo las mismas medias, 7.0 para el proceso A y 7.5 para el proceso B, pero ahora la misma diferencia de 0.5 entre los efectos observados de ambos procesos sería relativamente grande en comparación con el tamaño observado del error experimental. Esto sugeriría entonces que la diferencia observada entre ambos procesos sería en este caso atribuible a diferencias reales de eficacia entre ellos más que a su variabilidad inherente.

Por tanto, un **aspecto crucial en el diseño y el análisis de experimentos es que estos se planifiquen y realicen de manera que permitan aislar el efecto sobre la variable respuesta de las diferentes fuentes de variabilidad**. En particular, es importante entender que, si existen fuentes de variabilidad potencialmente elevada que no son identificadas, y el experimento no se planifica adecuadamente para poder aislarlas, su efecto se mezclará irremisiblemente con el del error experimental, lo que conllevará que este se *infle* o aumente de magnitud, y a su vez esto dificultará considerablemente que el efecto del factor de estudio sea relevante en comparación con este error experimental inflado.

Formalmente, estas comparaciones entre el efecto del factor de interés y el error experimental se llevan a cabo mediante una técnica estadística conocida como **análisis de la varianza o ANOVA** (por las siglas en inglés de *ANalysis Of VAriance*), que se basa en la noción de contraste o test de hipótesis. Específicamente, como se verá en la siguiente sección, el ANOVA toma la forma de un test en que se contrasta la hipótesis nula de igualdad de efectos entre los diferentes niveles del factor. Mediante este procedimiento es entonces posible imponer niveles de confianza (o de aversión al riesgo estadístico) con los que establecer **qué ratios entre el efecto del factor y la magnitud del error experimental son estadísticamente significativos** o, dicho de otro modo, cuándo la diferencia entre los efectos de los distintos niveles del factor es suficientemente grande en comparación con la magnitud estimada del error experimental para poder concluir, con el nivel de riesgo deseado, que el factor bajo estudio efectivamente tiene influencia sobre la respuesta de interés.

1.2. ANÁLISIS DE LA VARIANZA EN DISEÑOS UNIFACTORIALES

El análisis de la varianza (ANOVA) es un procedimiento estadístico, creado por R. A. Fisher en 1925, para descomponer la variabilidad de un experimento en componentes independientes

que pueden asociarse a causas distintas. El problema general que afronta el ANOVA se puede establecer como sigue: se dispone de *N* elementos y se observa en ellos una característica continua *Y* haciendo variar los niveles de un único factor *X*, en el que estamos interesados por su posible influencia sobre *Y*. La característica de interés *Y* se suele denominar *variable respuesta* o *respuesta* a secas. Sea *a* el número de niveles del factor *X* ($a \geq 2$). En este curso supondremos siempre que en cada nivel del factor *X* se observa la respuesta *Y* sobre un mismo número *n* de elementos, diferentes para cada nivel. Así pues, ha de ser $N = an$. El objetivo es conocer si el factor *X* tiene influencia sobre la respuesta *Y*, esto es, si existen diferencias entre los valores medios esperados de *Y* asociados a niveles diferentes de *X*. En otras palabras, podemos entender que se tienen *a* grupos de datos, cada uno con *n* observaciones, y el interés estriba en determinar si las medias de cada grupo son iguales o no.

Es importante subrayar que en este contexto el factor *X* se entiende como una variable explicativa categórica, que toma valores posiblemente no numéricos (como tipos de procesos de forjado en el ejemplo anterior) y cuya única función es dividir las observaciones de la respuesta en grupos asociados a cada uno de los niveles del factor. Además, supondremos que los niveles que toma el factor han sido elegidos y controlados por el experimentador, por lo que no se considera a *X* como una variable aleatoria, sino como un factor fijo. El modelo que surge bajo este supuesto se conoce como de **efectos fijos**, en contraposición al modelo de **efectos aleatorios** que surge al considerar que los niveles de *X* que intervienen en el experimento se muestrean aleatoriamente de una población más amplia de niveles.

Además, al considerar un único factor de interés en el experimento, este modelo se conoce como **unifactorial**. Es posible desarrollar modelos con más de un factor, así como utilizar una o más variables bloque en conjunción con cualquier número de factores. La complejidad de estos modelos sin embargo crece de manera importante al aumentar el número de factores y variables bloque.

Para poder generalizar los datos observados y trabajar abstractamente con ellos es necesario suponer un modelo matemático sobre el proceso que los genera. Así, denotaremos por $y_{ij} \in \mathbb{R}$ el valor observado de *Y* en el *j*-ésimo elemento asignado al nivel *i* del factor, con $j = 1,...,n$ e $i = 1,...,a$. Se asume que este valor observado oscila aleatoriamente alrededor de una media desconocida $\mu_i \in \mathbb{R}$, que representa el valor medio (poblacional) de las observaciones realizadas cuando *X* se fija en el nivel *i*. Se asume también que la variabilidad de estas observaciones y_{ij} con respecto al valor medio μ_i depende de una multitud de factores no controlados o no considerados, y llamaremos *error experimental* al término que engloba todos estos efectos no controlados. Así, denotando por ε_{ij} el error experimental o perturbación que interviene en la generación de la observación y_{ij}, se tiene el siguiente *modelo estadístico de las observaciones*, conocido como **modelo de las medias**:

$$y_{ij} = \mu_i + \varepsilon_{ij} \quad \forall i = 1,...,a, j = 1,...,n \tag{1.1}$$

Nótese que la perturbación $\varepsilon_{ij} = y_{ij} - \mu_i$ mide la desviación de la observación *j*-ésima realizada en el nivel *i* del factor con respecto a la media μ_i de *Y* en ese nivel. Téngase en cuenta también que, en tanto las medias μ_i son desconocidas, también han de serlo los errores ε_{ij}. En otras palabras, tanto las medias μ_i como los errores ε_{ij} son cantidades desconocidas e inobservables, cuya utilidad es proporcionar un modelo teórico de generación de los datos y_{ij}. No obstante, más adelante estudiaremos cómo estimar los parámetros μ_i de este modelo.

Una formulación equivalente de este modelo asume que las medias μ_i de los niveles se descomponen como $\mu_i = \mu + \tau_i \; \forall i = 1,...,a$, donde μ denota la media global de la variable *Y* y $\tau_i = \mu_i - \mu$ representa el efecto diferencial del nivel o grupo *i* respecto a esa media global. En otras palabras, se asume que

$$\frac{\sum_{i=1}^{a} \mu_i}{a} = \mu$$

lo que implica que se cumple la restricción $\sum_{i=1}^{a} \tau_i = 0$ en tanto los τ_i son desviaciones respecto a su media conjunta μ. Esto conduce al conocido como **modelo de los efectos**, dado por

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \;\; \forall i = 1,...,a, j = 1,...,n \tag{1.2}$$

que es inmediato comprobar es equivalente al anterior modelo de las medias. Nótese que tanto el modelo de las medias como el de los efectos son **modelos lineales**, en el sentido de que la variable respuesta y_{ij} se expresa como una función lineal de los parámetros μ_i o μ y τ_i. Obsérvese además que la diferencia entre las medias de dos niveles *i* e *i'* es igual a la diferencia entre los efectos diferenciales de esos niveles, esto es, $\mu_i - \mu_{i'} = \mu + \tau_i - (\mu + \tau_{i'}) = \tau_i - \tau_{i'}$.

1.2.1. Supuestos básicos del modelo

En este contexto, las perturbaciones o errores ε_{ij} representan la variabilidad intrínseca del experimento, aglomerando el efecto de otros factores con influencia en la respuesta pero no considerados en experimento, y poseen por tanto una naturaleza aleatoria, esto es, no predecible. Por tanto, los errores ε_{ij} son variables aleatorias, y supondremos que estas variables verifican los siguientes supuestos:

1. La esperanza de los errores es cero, esto es,

$$E[\varepsilon_{ij}] = 0 \;\; \forall i = 1,...,a, j = 1,...,n$$

2. La varianza de los errores es constante o, en otras palabras, la variabilidad respecto al valor medio μ_i es la misma en cada grupo *i*, *i* = 1,…,*a*. Esto se suele referir diciendo que los errores son *homocedásticos*. Formalmente,

$$V[\varepsilon_{ij}] = \sigma^2 \;\; \forall i = 1,...,a, \; j = 1,...,n$$

3. Los errores siguen una distribución normal. Atendiendo a los supuestos 1 y 2 anteriores, este supuesto se puede formalizar como sigue:

$$\varepsilon_{ij} \sim N(0, \sigma^2) \;\; \forall i = 1,...,a, \; j = 1,...,n$$

4. Los errores son independientes entre sí, por lo que la o las perturbaciones que influyen en la generación de unas observaciones no proporcionan información sobre cómo serán las perturbaciones asociadas a otras observaciones. Además, en tanto el supuesto 3 anterior establece que los errores son normales, y entre variables conjuntamente normales solo pueden existir relaciones lineales, este supuesto de independencia equivale a establecer que las covarianzas entre perturbaciones asociadas a observaciones diferentes son cero, esto es,

$$Cov[\varepsilon_{ij}, \varepsilon_{i'j'}] = 0 \text{ si } (i, j) \neq (i', j')$$

A este conjunto de 4 supuestos de corte probabilístico sobre los errores del modelo estadístico se los conoce como **supuestos básicos del modelo**. Merece la pena hacer un breve comentario sobre ellos y sobre la importancia de verificar su cumplimiento en la práctica, en tanto los procedimientos teóricos que se desarrollan más abajo dependen fuertemente de ellos, y las conclusiones que estos aportan pueden no ser válidas si estos supuestos no se cumplen.

El primer supuesto, que la media de los errores es 0, es una condición poco exigente, y que a la vez se puede justificar intuitivamente atendiendo a que los términos de error ε_{ij} acumulan la variabilidad resultante del conjunto de factores con influencia sobre la respuesta *Y* no considerados en el modelo. El efecto conjunto de estos factores producirá en algunos casos desviaciones positivas respecto a la media del nivel *i*, y negativas en otros, y la aleatorización del experimento conlleva que, a largo plazo, unas se compensarán con otras y el promedio de estas desviaciones será por tanto 0. En todo caso, si se supone que en algún nivel *i* el término de error tiene una esperanza $E[\varepsilon_{ij}] = s_i \neq 0$, es posible entonces definir una nueva perturbación $\rho_{ij} = \varepsilon_{ij} - s_i$, que ahora tendrá esperanza 0, de modo que la ecuación del modelo puede ser escrita como $y_{ij} = \mu_i + s_i + \rho_{ij}$, y renombrando $\mu_i' = \mu_i + s_i$ se tiene que las perturbaciones alrededor de esta nueva media tendrán, como se ha dicho, valor esperado 0.

El segundo supuesto es bastante más exigente, al suponer que la variabilidad de los errores es constante y estable. Sin embargo, es perfectamente posible que la dispersión de la respuesta en algunos grupos sea diferente a la de otros. Por tanto, este supuesto de errores homocedásticos debe comprobarse empíricamente, y veremos cómo hacerlo más adelante. En caso de que no se cumpla, es posible y habitual transformar la variable respuesta de modo que el modelo transformado sí tenga errores homocedásticos.

El tercer supuesto, la normalidad de los errores, es aproximadamente cierto en una gran variedad de situaciones en que la respuesta depende de otros muchos factores. De hecho, este es el significado profundo del Teorema Central de Límite. No obstante, este supuesto también ha de comprobarse empíricamente.

Finalmente, el cuarto supuesto, de independencia de los errores, implica que el orden de recogida de los datos, o equivalentemente de realización de los experimentos, no es informativo. Como con el primer supuesto, esto suele ser cierto en base a la aleatorización del experimento. En cualquier caso, también es posible comprobar empíricamente su verificación.

Una consecuencia directa de la consideración de estos supuestos básicos y de la introducción del modelo lineal es que la distribución de la respuesta *Y* dentro de cada nivel *i* sigue también una distribución normal, con media $\mu_i = \mu + \tau_i$ y varianza σ^2, esto es, $Y|_{X=i} \sim N(\mu_i, \sigma^2)$, o equivalentemente $y_{ij} \sim N(\mu_i, \sigma^2)\ \forall j = 1,...,n$. Otra consecuencia es que las observaciones y_{ij}, al igual que los errores, son también independientes entre sí, y en particular la covarianza de observaciones diferentes es nula, esto es, $Cov(y_{ij}, y_{i'j'}) = 0$ si $(i, j) \neq (i', j')$.

1.2.2. Estimación de los parámetros del modelo

Una vez introducido el modelo y los supuestos básicos en que se sustenta, podemos pasar a estudiar cómo estimar los parámetros μ_i del modelo de las medias (1.1). El procedimiento habitual para llevar a cabo esta estimación se conoce como el **criterio de mínimos cuadrados**. Este criterio se basa en obtener los valores de los parámetros μ_i, digamos $\hat{\mu}_i$, que minimizan la suma de al cuadrado de los errores $\varepsilon_{ij} = y_{ij} - \mu_i$, esto es, la función

$$L(\mu_1,...,\mu_a)=\sum_{i=1}^{a}\sum_{j=1}^{n}\varepsilon_{ij}^2=\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\mu_i)^2$$

Derivando esta función respecto a cada parámetro e igualando a cero se obtiene el siguiente sistema de *a* ecuaciones lineales, conocido como el sistema de **ecuaciones normales**:

$$\begin{aligned}\frac{\partial L(\mu_1,...,\mu_a)}{\partial \mu_1}&=-2\sum_{j=1}^{n}(y_{1j}-\mu_1)=0\\ &\vdots\\ \frac{\partial L(\mu_1,...,\mu_a)}{\partial \mu_a}&=-2\sum_{j=1}^{n}(y_{aj}-\mu_a)=0\end{aligned}$$

A partir de este sistema se obtienen las siguientes soluciones o estimadores de los parámetros:

$$\begin{aligned}\hat{\mu}_1&=\frac{1}{n}\sum_{j=1}^{n}y_{1j}\overset{\text{notación}}{=}\overline{y}_{1\cdot}\\ &\vdots\\ \hat{\mu}_a&=\frac{1}{n}\sum_{j=1}^{n}y_{aj}\overset{\text{notación}}{=}\overline{y}_{a\cdot}\end{aligned}$$

Así pues, el valor medio de la respuesta *Y* en el nivel *i* se estima mediante la media muestral de las observaciones de ese nivel, esto es, $\hat{\mu}_i=\overline{y}_{i\cdot}\ \forall i=1,...,a$.

La estimación de los parámetros μ y τ_i del modelo de los efectos (1.2) se realiza de manera similar. Sin embargo, en este caso se obtiene un sistema de *a* + 1 ecuaciones linealmente dependientes, con rango *a*, que no admite solución única. Esto se resuelve imponiendo la restricción

$$\sum_{i=1}^{a}\hat{\tau}_i=0$$

lo que conduce a los siguientes estimadores de los parámetros del modelo de los efectos (ejercicio):

$$\begin{aligned}\hat{\mu}&=\frac{1}{N}\sum_{i=1}^{a}\sum_{j=1}^{n}y_{ij}\overset{\text{notación}}{=}\overline{y}_{\cdot\cdot}\\ \hat{\tau}_1&=\frac{1}{n}\sum_{j=1}^{n}y_{1j}-\hat{\mu}=\overline{y}_{1\cdot}-\overline{y}_{\cdot\cdot}\\ &\vdots\\ \hat{\tau}_a&=\frac{1}{n}\sum_{j=1}^{n}y_{aj}-\hat{\mu}=\overline{y}_{a\cdot}-\overline{y}_{\cdot\cdot}\end{aligned}$$

Por tanto, en el modelo de los efectos el estimador de la media global de la respuesta *Y* viene dado por la media muestral de todas las observaciones, $\hat{\mu}=\overline{y}_{\cdot\cdot}$, mientras que el efecto diferencial de cada nivel *i* respecto a esta media global se estima como $\hat{\tau}_i=\overline{y}_{i\cdot}-\overline{y}_{\cdot\cdot}\ \forall i=1,...,a$, esto es, como la diferencia entre la media muestral de las observaciones del nivel *i* y la media muestral global.

Los supuestos básicos antes introducidos permiten obtener los momentos poblacionales y las distribuciones de estos estimadores, como se muestra en la siguiente proposición.

Proposición 1.1. Asumiendo los supuestos básicos, se tienen los siguientes resultados:

1) $\hat{\mu}_i = \overline{y}_{i\cdot} \sim N(\mu_i, \sigma^2/n)$

2) $\hat{\mu} = \overline{y}_{\cdot\cdot} \sim N(\mu, \sigma^2/N)$

3) $\hat{\tau}_i = \overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot} \sim N(\tau_i, \frac{N-n}{nN}\sigma^2)$

Demostración: La normalidad de los tres estimadores es consecuencia de que estos son combinación lineal de variables normales si se admiten los supuestos básicos. Veamos cómo obtener la esperanza y varianza de $\hat{\mu}_i = \overline{y}_{i\cdot}$:

$$E[\overline{y}_{i\cdot}] = E[\frac{1}{n}\sum_{j=1}^{n} y_{ij}] = \frac{1}{n}\sum_{j=1}^{n} E[y_{ij}] = \frac{n\mu_i}{n} = \mu_i$$

$$V[\overline{y}_{i\cdot}] = V[\frac{1}{n}\sum_{j=1}^{n} y_{ij}] = \frac{1}{n^2}\sum_{j,j'=1}^{n} Cov[y_{ij}, y_{ij'}] = \frac{1}{n^2}\sum_{j=1}^{n} V[y_{ij}] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

La obtención de los momentos de $\hat{\mu}$ y $\hat{\tau}_i$ se deja como ejercicio. □

A partir de las estimaciones anteriores de los parámetros es posible definir las nociones de **valores predichos** por el modelo lineal y sus **residuos**.

Definición 1.1. Se denominan *valores predichos* por el modelo lineal al resultado de sustituir, para cada observación, los parámetros en la expresión (1.1) o (1.2) por los correspondientes estimadores de mínimos cuadrados. Esto es, el valor predicho por el modelo para la observación $y_{ij}, i = 1,...,a, j = 1,...,n$ viene dado por

$$\hat{y}_{ij} = \hat{\mu}_i = \hat{\mu} + \hat{\tau}_i = \overline{y}_{i\cdot} \quad i = 1,...,a, j = 1,...,n$$

Así, cada observación y_{ij} en el nivel *i*, con *j* = 1,…,*n*, se aproxima por la media muestral de las observaciones de ese nivel.

Definición 1.2. Se denominan *residuos* del modelo a la diferencia entre los valores observados y los valores predichos, esto es,

$$e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \overline{y}_{i\cdot} \quad i = 1,...,a, \; j = 1,...,n$$

Podemos interpretar los residuos como estimaciones de los errores $\varepsilon_{ij} = y_{ij} - \mu_i$. Es importante observar que los residuos no son independientes, ya que están sometidos a ciertas restricciones derivadas de la estimación de los parámetros. En particular, sumando todos los residuos asociados al nivel *i* del factor, se obtiene

$$\sum_{j=1}^{n} e_{ij} = \sum_{j=1}^{n} (y_{ij} - \overline{y}_{i\cdot}) = \sum_{j=1}^{n} y_{ij} - n\overline{y}_{i\cdot} = 0 \qquad i = 1,...,a$$

Por tanto, se tienen *a* restricciones para los residuos, ya que su suma dentro de cada grupo o nivel ha de ser nula. En consecuencia, en cada grupo existirán *n* – 1 residuos no determinados a priori, estando el *n*-ésimo determinado a través de la restricción anterior. Sumando a lo largo de los *a* grupos, se obtiene que existirán en total *N* – *a* residuos no determinados, que pueden por tanto tomar cualquier valor. Por este motivo, se denomina **grados de libertad de los residuos** a la diferencia entre el número total de residuos, *N*, tantos como observaciones, y el número de restricciones lineales impuestas sobre ellos al estimar los parámetros, o equivalentemente al número de residuos no determinados a priori por estas restricciones. Así pues, los residuos tienen *N* – *a* grados de libertad.

1.2.3. Descomposición de la variabilidad en sumas de cuadrados

Como se adelantaba al comienzo de esta sección, el análisis de la varianza se fundamenta en la descomposición de la variabilidad de la respuesta en componentes independientes asociadas a causas diferentes. Veamos ahora cómo se formaliza esta descomposición.

Proposición 1.2. Se denomina **Suma de cuadrados total** (SCT) a la suma de los cuadrados de las desviaciones de las observaciones y_{ij} respecto de la media muestral global $\overline{y}_{..}$, esto es,

$$SCT = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \overline{y}_{..})^2 .$$

Además, se conoce como **Suma de cuadrados del factor** (SCF) a la suma de los cuadrados de las variaciones de las medias estimadas de los niveles $\hat{\mu}_i = \overline{y}_{i.}$ respecto de la media muestral global $\overline{y}_{..}$, es decir

$$SCF = n\sum_{i=1}^{a}(\overline{y}_{i.} - \overline{y}_{..})^2 .$$

Finalmente, se denomina **Suma de cuadrados de los residuos** (SCR) a la suma de los cuadrados de las desviaciones entre las observaciones y sus valores predichos, dada por

$$SCR = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \overline{y}_{i.})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}e_{ij}^{\ 2} .$$

Entonces, se cumple la igualdad ***SCT = SCF + SCR***, esto es,

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

Demostración: Para comenzar, nótese que $(y_{ij} - \overline{y}_{..}) = (\overline{y}_{i.} - \overline{y}_{..}) + (y_{ij} - \overline{y}_{i.})$. Elevando al cuadrado y sumando para todas las *N* observaciones, se tiene

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \overline{y}_{..})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}[(\overline{y}_{i.} - \overline{y}_{..}) + (y_{ij} - \overline{y}_{i.})]^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}[(\overline{y}_{i.} - \overline{y}_{..})^2 + (y_{ij} - \overline{y}_{i.})^2 + 2(\overline{y}_{i.} - \overline{y}_{..})(y_{ij} - \overline{y}_{i.})] =$$

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

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

Este último sumatorio con los productos cruzados se anula, pues se puede expresar como

$$2\sum_{i=1}^{a}(\overline{y}_{i.} - \overline{y}_{..})\sum_{j=1}^{n}(y_{ij} - \overline{y}_{i.}) = 2\sum_{i=1}^{a}(\overline{y}_{i.} - \overline{y}_{..})\sum_{j=1}^{n}e_{ij}$$

y ya se vio antes que $\sum_{j=1}^{n}e_{ij} = 0$ para todo *i*. Por tanto, se cumple la igualdad *SCT* = *SCF* + *SCR*. □

Cada una de estas sumas de cuadrados tiene una interpretación relevante:

- **SCT**, la suma de cuadrados total, representa la **variabilidad total** de las observaciones de la respuesta alrededor de su media global. En ausencia de un modelo lineal para cada nivel del factor como el dado por las expresiones (1.1) o (1.2), el modelo base para la respuesta *Y* sería el dado por $Y = \mu + \varepsilon$, que se estimaría mediante mínimos cuadrados como $\hat{\mu} = \overline{y}_{..}$ (ejercicio). Los residuos de este modelo serían las diferencias $y_{ij} - \overline{y}_{..}$, y la

suma de sus cuadrados SCT representa de algún modo la variabilidad de *Y* que no puede ser explicada simplemente por su media. Nótese además que las diferencias $y_{ij} - \overline{y}_{..}$ tienen $N - 1$ grados de libertad, ya que se cumple la restricción lineal $\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \overline{y}_{..}) = 0$ (ejercicio). Por esto, diremos que la SCT tiene $N - 1$ grados de libertad. Conviene observar también que la SCT dividida por $N - 1$, precisamente sus grados de libertad, es igual a la cuasivarianza muestral de *Y*, esto es,

$$\frac{SCT}{N-1} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \overline{y}_{..})}{N-1} = \frac{\sum_{l=1}^{N}(y_l - \overline{y})^2}{N-1} = s_Y^2$$

donde en el penúltimo término las observaciones de la respuesta no se consideran agrupadas para enfatizar que la SCT mide la variabilidad total de las observaciones en ausencia de un modelo explicativo como (1.1) o (1.2).

- **SCF**, la suma de cuadrados del factor, representa la variabilidad introducida en la respuesta por la aplicación en el experimento de diferentes niveles del factor, ya que se computa a partir de las diferencias $\overline{y}_{i\cdot} - \overline{y}_{..} = \hat{\mu}_i - \hat{\mu} = \hat{\tau}_i$ entre el nivel medio estimado de cada nivel y la media global de las observaciones, o equivalentemente a partir de los efectos diferenciales estimados de cada nivel. Por esto, la SCF crece cuando aumentan las diferencias entre las medias observadas en cada nivel. En este sentido, la SCF se puede interpretar como la **variabilidad explicada** por la introducción del modelo (1.1) o (1.2) que no podía ser explicada mediante la media global, y de hecho obsérvese que se obtiene a partir de las diferencias entre el valor predicho por el modelo (1.1) y (1.2) y el predicho por el anterior modelo con solo la media. Además, en tanto que solo se tienen a diferencias $\overline{y}_{i\cdot} - \overline{y}_{..} = \hat{\tau}_i$ y el mecanismo de estimación introducía la restricción $\sum_{i=1}^{a}\hat{\tau}_i = 0$, solo $a - 1$ de esas diferencias no están determinadas a priori, por lo que la SCF tiene $a - 1$ grados de libertad.

- **SCR**, la suma de cuadrados de los residuos, representa la **variabilidad residual o no explicada** por el modelo (1.1) o (1.2), en tanto se obtiene mediante las diferencias $y_{ij} - \overline{y}_{i\cdot} = y_{ij} - \hat{\mu}_i = y_{ij} - \hat{y}_{ij} = e_{ij}$ entre las observaciones y los valores predichos por este modelo, es decir, los residuos. Como ya se indicó anteriormente, solo hay $N - a$ residuos independientes, por lo que la SCR tiene $N - a$ grados de libertad.

De este modo, la descomposición en suma de cuadrados anterior puede interpretarse diciendo que **la variabilidad total de las observaciones de la respuesta *Y* se divide en dos partes**:

- Una parte de **variabilidad explicada** por las variaciones del factor *X* mediante el modelo lineal (1.1) o (1.2), en tanto este permite estimar una media en cada grupo o nivel en lugar de una única media. Este tipo de variabilidad también se conoce como variabilidad externa o entre niveles (*between-levels*), en tanto mide realmente la dispersión º las medias de los diferentes niveles.
- Y otra parte de **variabilidad no explicada** por las variaciones del factor *X*, la cual se acumula en los residuos. Aquí se incluye por tanto la variabilidad aleatoria comentada en la Sección 1.1, asociada a la multitud de factores que influyen en la respuesta y que no son controlados en el experimento. También se la denomina variabilidad interna o dentro de los niveles (*within-levels*), al medir la dispersión de las observaciones en cada nivel.

Al igual que la SCT dividida por sus grados de libertad producía un estadístico relevante, también al dividir la SCF y la SCR por sus correspondientes grados de libertad se obtienen unas cantidades, las conocidas como **medias de cuadrados**, que son muy relevantes para el ANOVA, como veremos un poco más abajo. Definamos entonces estas medias.

Definición 1.3. Se denomina **Media de cuadrados del factor** (MCF) al resultado de dividir la SCF por sus grados de libertad, esto es

$$MCF = \frac{SCF}{a-1} = \frac{n\sum_{i=1}^{a}(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2}{a-1}$$

Definición 1.4. Se denomina **Media de cuadrados de los residuos** (MCR) al resultado de dividir la SCR por sus grados de libertad, esto es,

$$MCR = \frac{SCR}{N-a} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \overline{y}_{i\cdot})^2}{N-a}$$

En el siguiente resultado se dan las esperanzas de estas medias de cuadrados.

Proposición 1.3. Las medias de cuadrados tienen las siguientes esperanzas:

1) $E[MCR] = \sigma^2$
2) $E[MCF] = \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1}$

Demostración:

1) $E[MCR] = E[\frac{SCR}{N-a}] = \frac{1}{N-a}E[\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \overline{y}_{i\cdot})^2] = \frac{1}{N-a}E[\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}^2 + \overline{y}_{i\cdot}^2 - 2y_{ij}\overline{y}_{i\cdot})] =$

$= \frac{1}{N-a}E[\sum_{i=1}^{a}\sum_{j=1}^{n}y_{ij}^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}\overline{y}_{i\cdot}^2 - 2\sum_{i=1}^{a}\sum_{j=1}^{n}y_{ij}\overline{y}_{i\cdot}] = \frac{1}{N-a}E[\sum_{i=1}^{a}\sum_{j=1}^{n}y_{ij}^2 + n\sum_{i=1}^{a}\overline{y}_{i\cdot}^2 - 2n\sum_{i=1}^{a}\overline{y}_{i\cdot}^2] =$

$= \frac{1}{N-a}E[\sum_{i=1}^{a}\sum_{j=1}^{n}y_{ij}^2 - n\sum_{i=1}^{a}\overline{y}_{i\cdot}^2] = \frac{1}{N-a}E[\sum_{i=1}^{a}\sum_{j=1}^{n}y_{ij}^2 - n\sum_{i=1}^{a}(\frac{1}{n}\sum_{j=1}^{n}y_{ij})^2] = \frac{1}{N-a}E[\sum_{i=1}^{a}\sum_{j=1}^{n}y_{ij}^2 - \frac{1}{n}\sum_{i=1}^{a}(\sum_{j=1}^{n}y_{ij})^2]$

Sustituyendo ahora la expresión del modelo (1.1) en y_{ij} se obtiene

$E[MCR] = \frac{1}{N-a}E[\sum_{i=1}^{a}\sum_{j=1}^{n}(\mu_i + \varepsilon_{ij})^2 - \frac{1}{n}\sum_{i=1}^{a}(\sum_{j=1}^{n}\mu_i + \varepsilon_{ij})^2] = \frac{1}{N-a}E[\sum_{i=1}^{a}\sum_{j=1}^{n}(\mu_i^2 + \varepsilon_{ij}^2 + 2\mu_i\varepsilon_{ij}) - \frac{1}{n}\sum_{i=1}^{a}(n\mu_i + \sum_{j=1}^{n}\varepsilon_{ij})^2 =$

$= \frac{1}{N-a}E[\sum_{i=1}^{a}\sum_{j=1}^{n}\mu_i^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}\varepsilon_{ij}^2 + 2\sum_{i=1}^{a}\sum_{j=1}^{n}\mu_i\varepsilon_{ij} - \frac{1}{n}(n^2\sum_{i=1}^{a}\mu_i^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}\varepsilon_{ij}^2 + 2n\sum_{i=1}^{a}\mu_i\sum_{j=1}^{n}\varepsilon_{ij})]$

Ahora se ha de usar la linealidad de la esperanza para introducirla en los sumatorios, y teniendo en cuenta que $E[\varepsilon_{ij}] = 0$ y $E[\varepsilon_{ij}^2] = \sigma^2$ (ejercicio), se obtiene

$E[MCR] = \frac{1}{N-a}(n\sum_{i=1}^{a}\mu_i^2 + N\sigma^2 - n\sum_{i=1}^{a}\mu_i^2 - a\sigma^2) = \frac{N-a}{N-a}\sigma^2 = \sigma^2$

2) Ejercicio (usar modelo (1.2) en lugar de (1.1) al sustituir en y_{ij}). □

Así pues, la MCR proporciona **un estimador insesgado de** σ^2, la varianza supuesta constante y estable entre niveles del error experimental. Como se indicaba al final de la Sección 1.1, el procedimiento del análisis de la varianza precisa de una estimación de la magnitud del error experimental para analizar si la variabilidad producida por las variaciones en los niveles del factor es lo suficientemente grande como para poder asegurar con cierta confianza que tales

variaciones observadas en la respuesta no son debidas simplemente al error experimental. Pues bien, esa estimación vendrá dada precisamente por la MCR. Del mismo modo, la MCF proporciona una medida de la magnitud de la variabilidad de la respuesta debida a las variaciones del factor, que se comparará con la MCR de cara a analizar la influencia del factor.

Por otro lado, es importante observar que si no hay diferencias en la influencia sobre la respuesta de los diferentes niveles del factor, entonces todas las medias de los niveles serán iguales, i.e. $\mu_1 = ... = \mu_a = \mu$, o equivalentemente todos los efectos diferenciales de los niveles serán nulos, i.e. $\tau_1 = ... = \tau_a = 0$. Si este es el caso, nótese que entonces la esperanza de la MCF es también igual a σ^2, pues el término $\sum_{i=1}^{a} \tau_i^2$ se anula si $\tau_1 = ... = \tau_a = 0$. Por tanto, cuando el factor *X* no tiene influencia sobre la respuesta *Y*, tanto la MCR como la MCF proporcionarán estimadores insesgados de la varianza σ^2 del error experimental. En esta situación, ambas medias de cuadrados tomarán valores similares, o equivalentemente el ratio

$$F_0 = \frac{MCF}{MCR}$$

tomará valores próximos a 1. Si el factor *X* efectivamente tiene influencia sobre la respuesta, entonces los efectos τ_i de algunos niveles serán distintos de 0, y el término $\sum_{i=1}^{a} \tau_i^2$ será positivo. En este caso, la MCF tenderá a ser mayor que la MCR, y el ratio F_0 anterior tenderá a ser mayor que 1.

Este razonamiento proporciona ya una idea intuitiva de cómo analizar si el factor *X* tiene influencia sobre la respuesta a través del ratio F_0. Sin embargo, aún no podemos establecer qué valor de este ratio F_0 marca el punto de corte a partir del cual considerar la diferencia entre la MCR y la MCF como *suficientemente grande* para concluir que el factor realmente influye en la respuesta con cierta confianza. Para ello, necesitamos conocer la distribución de probabilidad del estadístico F_0, para lo que a su vez es necesario conocer las distribuciones de las medias de cuadrados MCF y MCR o, más específicamente, de las respectivas sumas de cuadrados SCF y SCR. Este es el objeto del teorema de Cochran, que demostramos en la siguiente sección.

1.2.4. Teorema de Cochran

Como acabamos de explicar, para poder llevar a un marco de estadística inferencial la comparación entre las medias de cuadrados MCF y MCR a través del ratio F_0 necesitamos conocer la distribución de este ratio, que a su vez depende de las distribuciones de las sumas de cuadrados SCF y SCR en que se basan las medias MCF y MCR. Intuitivamente, sabemos que la distribución de una suma de cuadrados de variables normales está relacionada con la distribución chi-cuadrado. Pero necesitamos un resultado general que proporcione esta conclusión, nos informe de los grados de libertad de cada suma de cuadrados y, no menos importante, nos garantice la independencia de estas, sin la cual no podremos establecer la distribución del ratio F_0. Este resultado lo proporciona un teorema atribuido a William G. Cochran, uno de los continuadores de la labor de R.A. Fisher. Para poder demostrarlo, necesitaremos recordar algunos resultados básicos de álgebra lineal y probar algunos otros resultados sobre distribuciones de formas cuadráticas de variables normales.

Lema 1.1. Sea $\mathbf{Y} = (Y_1, ..., Y_N) \in \mathbb{R}^N$ y sea **X** una matriz *N* x *p* cuyas columnas son una base de cierto subespacio vectorial $E_p \subset \mathbb{R}^N$. Entonces la proyección ortogonal de **Y** sobre el espacio E_p viene

dada por **AY**, donde la matriz cuadrada **A** es simétrica, idempotente, de rango *p* y tal que $\mathbf{A}=\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$.

Demostración: Si $\mathbf{V}=\mathbf{AY}$ es la proyección ortogonal de **Y** sobre un cierto subespacio E_p, generado por las columnas de **X**, entonces la proyección de **V** sobre ese mismo subespacio debe ser invariante, al estar ya **V** en E_p. Por tanto, ha de ser **AV** = **V**, lo que implica que **A**(**AY**) = **AY** para todo **Y**, y por tanto se tiene que cumplir que **A** = $\mathbf{A}^2$, es decir, la matriz proyección **A** debe ser idempotente. Veamos que esta matriz **A** ha de venir dada por $\mathbf{A}=\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$. Claramente, la matriz **A** así definida es simétrica e idempotente. Veamos entonces que **A** no depende de la base elegida. Si consideramos otra base **B** dada por **B** = **XC**, donde **C** es *p* x *p* y no singular, se tendrá que, como $(\mathbf{MNL})^{-1}=\mathbf{L}^{-1}\mathbf{N}^{-1}\mathbf{M}^{-1}$ para matrices cuadradas no singulares, se cumple que $\mathbf{B}(\mathbf{B}^t\mathbf{B})^{-1}\mathbf{B}^t=\mathbf{XC}(\mathbf{C}^t\mathbf{X}^t\mathbf{XC})^{-1}\mathbf{C}^t\mathbf{X}^t=\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t=\mathbf{A}$, por lo que **A** no depende de la base escogida. Veamos ahora que el vector **V** definido por **V** = **AY** verifica las condiciones de una proyección ortogonal sobre el espacio $E_p=span(\mathbf{X})$ generado por las columnas de **X**. En primer lugar, demostremos que **V** está contenido en E_p. Sea $\boldsymbol{\beta}=(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$ y denotemos por $\mathbf{X}_1,\ldots,\mathbf{X}_p$ las columnas de **X**. Entonces $\mathbf{V}=\mathbf{X}\boldsymbol{\beta}=\beta_1\mathbf{X}_1+\ldots+\beta_p\mathbf{X}_p$, por lo que **V** ha de pertenecer a E_p al ser combinación lineal de las columnas de **X**. Veamos ahora que **Y – V** es ortogonal a E_p. Todo vector $\mathbf{U}\in E_p$ se puede expresar como $\mathbf{U}=\alpha_1\mathbf{X}_1+\ldots+\alpha_p\mathbf{X}_p=\mathbf{X}\boldsymbol{\alpha}$, y por tanto $\mathbf{U}^t(\mathbf{Y}-\mathbf{V})=\mathbf{U}^t(\mathbf{I}-\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t)\mathbf{Y}=\boldsymbol{\alpha}^t(\mathbf{X}^t-\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t)\mathbf{Y}=\mathbf{0}$, por lo que **Y** – **V** es ortogonal a cualquier vector de E_p, lo que demuestra el resultado. □

Lema 1.2. Sea **A** una matriz cuadrada *N* x *N*. **A** es la matriz de proyección ortogonal sobre un cierto subespacio E_p si y solo si **A** es simétrica idempotente de rango *p*.

Demostración: Si **A** define una proyección ortogonal, por el Lema 1.1 se tiene que $\mathbf{A}=\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$, con **X** una matriz *N* x *p* cuyas columnas forman una base del espacio sobre el que se proyecta, y por tanto **A** es simétrica e idempotente y tiene rango *p*. Por otro lado, supongamos ahora que **A** es simétrica idempotente de rango *p*, y sea $\mathbf{Y}\in\mathbb{R}^N$ un vector cualquiera, que puede expresarse como **Y** = **AY** + (**I** - **A**)**Y**. Veamos que (**I** - **A**)**Y** es ortogonal a todo vector que pertenezca al subespacio $E_p=\{\mathbf{AC}\,|\,\mathbf{C}\in\mathbb{R}^N\}$. Sea entonces **AC** un vector de E_p. Se tiene que $(\mathbf{AC})^t(\mathbf{I}-\mathbf{A})\mathbf{Y}=\mathbf{C}^t(\mathbf{A}^t-\mathbf{A}^t\mathbf{A})\mathbf{Y}=\mathbf{C}^t(\mathbf{A}-\mathbf{A})\mathbf{Y}=0$, por lo que **AY** es la proyección ortogonal sobre E_p. □

Lema 1.3. Sea **Y** = (Y_1,…, Y_N) un vector de *N* variables aleatorias independientes, con distribución conjunta $\mathbf{Y}\sim N(\boldsymbol{\mu},\sigma^2\mathbf{I})$, y sea **C** una matriz *N* x *N* ortogonal (esto es, tal que $\mathbf{C}^t\mathbf{C}=\mathbf{C}\mathbf{C}^t=\mathbf{I}$). Entonces el vector **Z** = (Z_1,…, Z_N) obtenido como **Z** = **CY** cumple que $\mathbf{Z}\sim N(\mathbf{C}\boldsymbol{\mu},\sigma^2\mathbf{I})$.

Demostración: El vector **Z** se distribuye normalmente al obtenerse como combinación lineal de variables conjuntamente normales. Además, se tiene que $E[\mathbf{Z}]=E[\mathbf{CY}]=\mathbf{C}E[\mathbf{Y}]=\mathbf{C}\boldsymbol{\mu}$ y $V[\mathbf{Z}]=E[(\mathbf{Z}-E[\mathbf{Z}])(\mathbf{Z}-E[\mathbf{Z}])^t]=E[(\mathbf{CY}-\mathbf{C}\boldsymbol{\mu})(\mathbf{CY}-\mathbf{C}\boldsymbol{\mu})^t]=\mathbf{C}E[(\mathbf{Y}-\boldsymbol{\mu})(\mathbf{Y}-\boldsymbol{\mu})^t]\mathbf{C}^t=\mathbf{C}V[\mathbf{Y}]\mathbf{C}^t=\mathbf{C}\sigma^2\mathbf{I}\mathbf{C}^t=$ $=\sigma^2\mathbf{I}$. □

Lema 1.4. Sea **Y** un vector de *N* variables normales estándar e independientes, esto es, $\mathbf{Y}\sim N(\mathbf{0},\mathbf{I})$, y sea **A** una matriz cuadrada *N* x *N*, con la que se define la forma cuadrática *S* = $\mathbf{Y}^t\mathbf{AY}$. Si **A** es simétrica idempotente de rango *r*, entonces *S* se distribuye como una chi-cuadrado con *r* grados de libertad, esto es, $S\sim\chi_r^2$.

Demostración: Como **A** es simétrica, existe una matriz ortogonal **C** que la diagonaliza. Sea **X** = **CY**. Por el Lema 1.3, el vector **X** tendrá una distribución $\mathbf{X}\sim N(\mathbf{0},\mathbf{I})$. Además,

$S = \mathbf{Y}^t\mathbf{AY} = \mathbf{X}^t\mathbf{CAC}^t\mathbf{X} = \mathbf{X}^t\mathbf{DX}$, donde **D** es diagonal, y por ser **A** idempotente de rango *r* se cumple que **D** ha de tener *r* unos en la diagonal principal y el resto ceros. Por tanto, se tiene que

$$\mathbf{Y}^t\mathbf{AY} = \mathbf{X}^t\mathbf{DX} = \sum_{i=1}^{r} X_i^2 \sim \chi_r^2 \qquad \square$$

Corolario 1.1. Sea $\mathbf{Y} \sim N(\mathbf{0},\mathbf{I})$ un vector de *N* variables normales estándar e independientes. Entonces el cuadrado del módulo de su proyección sobre un espacio de dimensión *r* se distribuye como una chi-cuadrado con *r* grados de libertad.

Demostración: La proyección de **Y** sobre un subespacio vectorial E_r vendrá dada por **AY**, donde **A** es simétrica e idempotente de rango *r*. Entonces, el cuadrado del módulo de esta proyección será $(\mathbf{AY})^t(\mathbf{AY}) = \mathbf{Y}^t\mathbf{AY}$, y se aplica el anterior Lema 1.4 para obtener la conclusión. □

Lema 1.5. Sea $\mathbf{Y} \sim N(\mathbf{0},\mathbf{I})$ un vector de *N* variables normales estándar e independientes, y sean **A** y **B** dos matrices simétricas e idempotentes, que definen respectivamente las formas cuadráticas $S = \mathbf{Y}^t\mathbf{AY}$ y $T = \mathbf{Y}^t\mathbf{BY}$. En estas condiciones, se tiene que si $\mathbf{AB} = \mathbf{0}$ entonces *S* y *T* son independientes.

Demostración: Según el Corolario 1.1 toda forma cuadrática con matriz simétrica e idempotente puede expresarse como el cuadrado de módulo de un vector. Para la forma cuadrática *S* este vector será **AY**, y para *T* será **BY**. Si demostramos que estos vectores **AY** y **BY** son independientes, también lo serán los cuadrados de sus módulos, y por tanto las formas cuadráticas *S* y *T*. Para ello, veremos que ambos vectores están incorrelados, por lo que, al tener distribución conjunta normal multivariante, serán independientes. La matriz de covarianzas entre estos vectores es $E[(\mathbf{AY})(\mathbf{BY})^t] = E[\mathbf{AYY}^t\mathbf{B}^t] = \mathbf{A}E[\mathbf{YY}^t]\mathbf{B}^t = \mathbf{AB}^t = \mathbf{AB} = \mathbf{0}$, y por tanto los vectores **AY** y **BY** son independientes, lo que demuestra el resultado. □

Ya estamos en condiciones de enunciar y demostrar el teorema de Cochran.

Teorema de Cochran. Sea $\mathbf{Y} \sim N(\mathbf{0},\mathbf{I})$ un vector de *N* variables normales estándar e independientes. Supongamos que **Y** se puede descomponer como la suma de *h* vectores $\mathbf{S}_1,\ldots,\mathbf{S}_h$, i.e.

$$\mathbf{Y} = \mathbf{S}_1 + \ldots + \mathbf{S}_h$$

donde cada vector $\mathbf{S}_j$, $j = 1,\ldots,h$, es la proyección ortogonal de **Y** sobre un subespacio vectorial $E_j \subset \mathbb{R}^N$, y estos subespacios son ortogonales entre sí dos a dos, i.e. $\mathbf{S}_j^t\mathbf{S}_k = 0$ para todo $j \neq k$. Entonces el cuadrado del módulo de cada vector $\mathbf{S}_j$ (i.e. $|\mathbf{S}_j|^2 = \mathbf{S}_j^t\mathbf{S}_j$) se distribuye como una chi-cuadrado con un número de grados de libertad igual a la dimensión del subespacio E_j al que pertenece $\mathbf{S}_j$, i.e. $\mathbf{S}_j^t\mathbf{S}_j \sim \chi^2_{\dim(E_j)}$, $j = 1,\ldots,h$, y estas distribuciones son independientes dos a dos. Además, $\sum_{j=1}^{h} \dim(E_j) = N$.

Demostración: Si cada vector $\mathbf{S}_i$, $i = 1,\ldots,h$, resulta de la proyección ortogonal de **Y** sobre un subespacio vectorial, por el Lema 1.2 estos podrán expresarse en la forma $\mathbf{S}_i = \mathbf{A}_i\mathbf{Y}$, donde $\mathbf{A}_i$ es una matriz simétrica e idempotente de rango igual a la dimensión del espacio E_i que contiene a $\mathbf{S}_i$, esto es, $rango(\mathbf{A}_i) = \dim(E_i)$. Nótese que entonces el cuadrado del módulo de cada vector $\mathbf{S}_i$ viene a su vez dado por la forma cuadrática $\mathbf{S}_i^t\mathbf{S}_i = (\mathbf{A}_i\mathbf{Y})^t(\mathbf{A}_i\mathbf{Y}) = \mathbf{Y}^t\mathbf{A}_i\mathbf{A}_i\mathbf{Y} = \mathbf{Y}^t\mathbf{A}_i\mathbf{Y}$, y por el Lema 1.4 cada forma cuadrática $\mathbf{S}_i^t\mathbf{S}_i$ se distribuirá como una chi-cuadrado con $rango(\mathbf{A}_i) = \dim(E_i)$ grados de libertad, $i = 1,\ldots,h$. Además, si cada par de vectores $\mathbf{S}_i,\mathbf{S}_j$ con $i \neq j$ son ortogonales entre

sí, se tiene que $\mathbf{S}_i^t\mathbf{S}_j = \mathbf{Y}^t\mathbf{A}_i\mathbf{A}_j\mathbf{Y} = 0$ lo que implica que $\mathbf{A}_i\mathbf{A}_j = 0$ para todo $i \neq j$, y por el Lema 1.5 las formas cuadráticas anteriores son independientes entre sí. Finalmente, obsérvese que se cumple la igualdad $\mathbf{Y}^t\mathbf{Y} = \mathbf{Y}^t\mathbf{A}_1\mathbf{Y} + \mathbf{Y}^t\mathbf{A}_2\mathbf{Y} + ... + \mathbf{Y}^t\mathbf{A}_h\mathbf{Y}$ (teorema de Pitágoras), que implica que $\mathbf{A}_1 + \mathbf{A}_2 + ... + \mathbf{A}_h = \mathbf{I}$, y por la ortogonalidad de estas matrices ha de ser $\sum_{j=1}^{h}\dim(E_j) = \sum_{j=1}^{h} rango(\mathbf{A}_j) = N$. □

Así pues, la clave para la aplicación de este teorema es la descomposición de un vector $\mathbf{Y} \sim N(\mathbf{0},\mathbf{I})$ en una suma de componentes ortogonales entre sí. ¿Cómo se relaciona esto con las sumas de cuadrados SCF y SCR y su distribución de probabilidad? La clave está en la siguiente identidad, válida para cada observación y_{ij}, i =1,…,a, j = 1,…,n:

$$\frac{1}{\sigma}(y_{ij} - \mu) = \frac{1}{\sigma}(\overline{y}_{..} - \mu) + \frac{1}{\sigma}(\overline{y}_{i.} - \overline{y}_{..}) + \frac{1}{\sigma}(y_{ij} - \overline{y}_{i.})$$

Según hemos visto, la introducción del modelo lineal (1.1) y de los supuestos básicos conlleva que la variable respuesta Y sigue una distribución $N(\mu_i, \sigma^2)$ en cada grupo i asociado a un nivel del factor X, es decir, $y_{ij} \sim N(\mu_i, \sigma^2)\ \forall j = 1,...,n$. Como se discutió más arriba, cuando este factor no tiene influencia sobre la respuesta, todas las medias condicionadas serán iguales, esto es, $\mu_1 = ... = \mu_a = \mu$, y entonces las variables aleatorias $(y_{ij} - \mu)/\sigma$ en el término de la izquierda de la identidad anterior seguirán una distribución $N(0,1)$.

Por otro lado, es posible expresar vectorialmente el conjunto de las identidades anteriores para todas las observaciones mediante los siguientes vectores de dimensión $N = an$:

$$\mathbf{Y}^t = \frac{1}{\sigma}(y_{11} - \mu, y_{12} - \mu, ..., y_{1n} - \mu, y_{21} - \mu, ..., y_{an} - \mu)$$

$$\mathbf{S}_1^t = \frac{1}{\sigma}(\overline{y}_{..} - \mu, ..., \overline{y}_{..} - \mu)$$

$$\mathbf{S}_2^t = \frac{1}{\sigma}(\overline{y}_{1.} - \overline{y}_{..}, ..., \overline{y}_{2.} - \overline{y}_{..}, ..., \overline{y}_{a.} - \overline{y}_{..})$$

$$\mathbf{S}_3^t = \frac{1}{\sigma}(y_{11} - \overline{y}_{1.}, y_{12} - \overline{y}_{1.}, ..., y_{an} - \overline{y}_{a.})$$

Así, la anterior identidad se puede escribir para las $N = an$ observaciones como $\mathbf{Y} = \mathbf{S}_1 + \mathbf{S}_2 + \mathbf{S}_3$, y se cumple que $\mathbf{Y} \sim N(\mathbf{0},\mathbf{I})$ y que los vectores $\mathbf{S}_1$, $\mathbf{S}_2$ y $\mathbf{S}_3$ son ortogonales entre sí, puesto que

$$\mathbf{S}_1^t\mathbf{S}_2 = \sum_{i=1}^{a}\sum_{j=1}^{n}(\overline{y}_{..} - \mu)(\overline{y}_{i.} - \overline{y}_{..}) = 0$$

$$\mathbf{S}_1^t\mathbf{S}_3 = \sum_{i=1}^{a}\sum_{j=1}^{n}(\overline{y}_{..} - \mu)(y_{ij} - \overline{y}_{i.}) = 0$$

$$\mathbf{S}_2^t\mathbf{S}_3 = \sum_{i=1}^{a}\sum_{j=1}^{n}(\overline{y}_{i.} - \overline{y}_{..})(y_{ij} - \overline{y}_{i.}) = 0$$

Veamos ahora la dimensión de los subespacios ortogonales a los que pertenecen los diferentes vectores. Por las consecuencias de los supuestos básicos, el vector $\mathbf{Y}$ tiene N componentes independientes, y por tanto se mueve en un espacio de dimensión N. El vector $\mathbf{S}_1$ tiene todas sus N componentes iguales, y pertenecerá por tanto a un subespacio de dimensión 1. El vector

$\mathbf{S}_2$ tiene *a* componentes diferentes, aunque al cumplirse la restricción lineal $\sum_{i=1}^{a}(\overline{y}_{i\cdot}-\overline{y}_{\cdot\cdot})=0$ pertenecerá a un subespacio de dimensión *a* – 1. Finalmente, el vector $\mathbf{S}_3$ tiene *N* componentes diferentes, pero sujetas a las *a* restricciones $\sum_{j=1}^{n}(y_{ij}-\overline{y}_{i\cdot})=0$, *i* = 1,…,*a*, por lo que se moverá en un espacio de dimensión *N* – *a*. Nótese que *N* = 1 + (*a* – 1) + (*N* – *a*), por lo que la suma de los tres subespacios ortogonales asociados a los vectores $\mathbf{S}_1$, $\mathbf{S}_2$ y $\mathbf{S}_3$ es el espacio vectorial $\mathbb{R}^N$ en que se mueve el vector de observaciones.

En este contexto, la descomposición en sumas de cuadrados vista en la Sección 1.2.3 surge naturalmente como una aplicación del teorema de Pitágoras al triángulo formado por los vectores $\mathbf{Y}-\mathbf{S}_1$, $\mathbf{S}_2$ y $\mathbf{S}_3$, en tanto se cumple que $\mathbf{Y}-\mathbf{S}_1=\mathbf{S}_2+\mathbf{S}_3$, con $\mathbf{S}_2$ y $\mathbf{S}_3$ ortogonales, por lo que ha de ser $|\mathbf{Y}-\mathbf{S}_1|^2=|\mathbf{S}_2|^2+|\mathbf{S}_3|^2$. Esta expresión es de hecho la descomposición en sumas de cuadrados SCT = SCF + SCR, ya que $\text{SCT}/\sigma^2=|\mathbf{Y}-\mathbf{S}_1|^2$, $\text{SCF}/\sigma^2=|\mathbf{S}_2|^2$ y $\text{SCR}/\sigma^2=|\mathbf{S}_3|^2$. Y obsérvese que los grados de libertad discutidos en la Sección 1.2.3 para cada suma de cuadrados coinciden con las dimensiones de los respectivos subespacios vectoriales recién expuestas.

Para terminar, es claro por la discusión anterior que, al suponer que el factor *X* no tiene influencia sobre la respuesta y por tanto que todas las medias μ_i de la respuesta en los *a* niveles del factor *X* son iguales, se cumplen las condiciones para aplicar el teorema de Cochran a la descomposición $\mathbf{Y}=\mathbf{S}_1+\mathbf{S}_2+\mathbf{S}_3$. Esto lleva a concluir que los cuadrados de los módulos de los vectores $\mathbf{S}_1$, $\mathbf{S}_2$ y $\mathbf{S}_3$ siguen distribuciones chi-cuadrado con 1, *a* – 1 y *N* – *a* grados de libertad, respectivamente, y además que estas distribuciones son independientes entre sí. Esto es, hemos demostrado que, si $\mu_1=\ldots=\mu_a=\mu$ entonces **las sumas de cuadrados del factor, SCF, y residual, SCR, son independientes y se distribuyen como**

$$\frac{\text{SCF}}{\sigma^2}\sim\chi^2_{a-1} \quad \text{y} \quad \frac{\text{SCR}}{\sigma^2}\sim\chi^2_{N-a}$$

Si por el contrario el factor *X* sí tiene influencia sobre la respuesta y no todas las medias μ_i son iguales, entonces la suma de cuadrados SCF ya no se distribuye como una χ^2_{a-1} . Sin embargo, la **SCR siempre se distribuye como una** χ^2_{N-a} , independientemente de si el factor *X* tiene influencia sobre la respuesta *Y* o no (ejercicio).

1.2.5. Contraste de análisis de la varianza

Ya tenemos ahora todos los ingredientes listos para plantear rigurosamente el contraste ANOVA con el que dilucidar con cierta confianza si el factor *X* tiene o no influencia sobre la respuesta *Y* en base a las observaciones realizadas.

Como hemos visto, suponer que el factor no tiene efecto sobre la respuesta es equivalente a considerar que la media μ_i de la respuesta en cada nivel del factor es la misma. En tanto el objetivo del experimento es demostrar empíricamente la influencia del factor sobre la respuesta, el supuesto de partida del experimento y de su análisis debería ser que esta influencia no existe, y solo abandonar este supuesto cuando la evidencia disponible apunte claramente a que las diferencias observadas en la respuesta en los diferentes niveles solo son explicables si existe un efecto real del factor. En otras palabras, la **hipótesis nula del contraste ANOVA** será que todos los *a* grupos asociados a niveles del factor son iguales, esto es,

$$H_0:\mu_1=\ldots=\mu_a=\mu$$

y la **hipótesis alternativa** se obtiene como su negación, y afirma que al menos uno de los *a* grupos es diferente al resto, esto es,

$$H_1 : \mu_i \neq \mu \text{ para algún } i = 1,...,a$$

Estas hipótesis pueden plantearse de manera equivalente en términos de los efectos de los niveles del factor en lugar de en términos de sus medias:

$$H_0 : \tau_1 = ... = \tau_a = 0$$

$$H_1 : \tau_i \neq 0 \text{ para algún } i = 1,...,a$$

Si se supone que la hipótesis nula H_0 es cierta, hemos visto que las sumas de cuadrados SCF y SCR divididas por σ^2 siguen distribuciones chi-cuadrado independientes, por lo que al dividirlas por sus grados de libertad y realizar su cociente se obtiene un estadístico con distribución *F* de Fisher, esto es,

$$\frac{\frac{1}{a-1}\frac{SCF}{\sigma^2}}{\frac{1}{N-a}\frac{SCR}{\sigma^2}} = \frac{\frac{SCF}{a-1}}{\frac{SCR}{N-a}} = \frac{MCF}{MCR} = F_0 \sim F_{a-1,N-a}$$

En otras palabras, bajo H_0 el ratio F_0 entre la variabilidad explicada por el factor y la variabilidad debida al error experimental sigue una distribución *F* con $a - 1$ grados de libertad en el numerador y $N - a$ grados de libertad en el denominador. Cuando H_0 no es cierta, el denominador de F_0 seguirá la misma distribución, pero el numerador será en promedio mayor que una chi-cuadrado, por lo que tenderá a producirse un valor de F_0 *inusualmente alto* para una distribución *F*.

Como sabemos la distribución de F_0 bajo la hipótesis nula, ahora tenemos una herramienta para establecer con precisión qué significa eso de *inusualmente alto*. Denotemos por $F_{a-1,N-a;1-\alpha}$ el percentil $100(1-\alpha)\%$ de la distribución $F_{a-1,N-a}$, esto es, un valor tal que $P(F_{a-1,N-a} > F_{a-1,N-a;1-\alpha}) = \alpha$, donde $\alpha \in (0,1)$ se conoce como el *nivel de significación*. Entonces, solo en un $100\alpha\%$ de los casos (es decir, experimentos) se obtendrá un valor de F_0 mayor que $F_{a-1,N-a;1-\alpha}$ cuando H_0 es cierta. Por tanto, para valores de α relativamente pequeños, esto es, α = 0.05 o menor, obtener un valor de F_0 mayor que $F_{a-1,N-a;1-\alpha}$ puede tener dos explicaciones:

1) El experimento ha producido unas observaciones de la respuesta relativamente muy poco probables si el factor no tuviera influencia, esto es, si H_0 fuera cierta.
2) El factor tiene influencia sobre la respuesta, esto es, H_0 es falsa, en cuyo caso los valores observados de la respuesta serían mucho más probables.

En tanto la explicación 1) es mucho menos probable que la explicación 2) dada la evidencia disponible, cuando se obtiene un valor de F_0 mayor que $F_{a-1,N-a;1-\alpha}$ se entiende que la evidencia apunta a que se ha de rechazar la hipótesis nula H_0, por lo que se concluirá que el factor *X* tiene influencia sobre la respuesta *Y*. En caso contrario, si $F_{a-1,N-a} \leq F_{a-1,N-a;1-\alpha}$ se interpreta que la evidencia disponible no es suficiente para descartar la hipótesis nula H_0, por lo que no se puede concluir que el factor *X* sea influyente sobre la respuesta.

El procedimiento operativo para la realización del contraste ANOVA puede resumirse entonces en los siguientes pasos:

1. Decidir el nivel de significación α.
2. Obtener el estadístico $F_0 = \frac{MCF}{MCR}$.
3. Si $F_{a-1,N-a} > F_{a-1,N-a;1-\alpha}$ rechazar la hipótesis nula H_0. En caso contrario, no se puede descartar H_0.

Siempre que se realiza un contraste de hipótesis es conveniente obtener el nivel crítico del test o **p-valor**, que es el mínimo nivel de significación α que conduciría a un rechazo de H_0. Es importante entender que conocer el p-valor es más informativo que solamente conocer si el test rechaza H_0 o no, porque el p-valor indica además el **grado en que la evidencia soporta el rechazo de la hipótesis nula**. Cuanto más pequeño es el p-valor, mayor es la evidencia de que la hipótesis nula debe ser rechazada. Comunicar el resultado del test solo en términos de rechazo o no de H_0 obliga además a las personas que reciban esa información a trabajar con el mismo nivel de significación que condujo a ese resultado, mientras que comunicar el p-valor permite valorar el nivel de evidencia disponible y tomar una decisión en función del riesgo que cada cual esté dispuesto a asumir.

Operativamente, el p-valor del contraste ANOVA se calcula como la probabilidad de que el estadístico F_0 tome un valor mayor al observado bajo la hipótesis nula, esto es,

$$p-valor = P(F_{a-1,N-a} > F_0)$$

En la práctica, los paquetes de software estadístico suelen presentar la información relativa al contraste ANOVA en un formato de tabla estandarizado, que se conoce como **tabla ANOVA**, y que se muestra en la Figura 1.2.

FUENTE DE VARIABILIDAD	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	CUADRADOS MEDIOS	F_0	p-valor
FACTOR	SCF	$a-1$	MCF	MCF/MCR	$P(F_{a-1,N-a} > F_0)$
ERROR	SCR	$N-a$	MCR		
TOTAL	SCT	$N-1$			

Figura 1.2: Tabla ANOVA.

Aparte del p-valor, una medida relativa de la bondad del ajuste de los datos proporcionado por el modelo (1.1) o (1.2) viene dada por el denominado ***coeficiente de determinación***, o R^2, dado por

$$R^2 = \frac{SCF}{SCT} = 1 - \frac{SCR}{SCT}$$

Nótese que, por la descomposición en sumas de cuadrados, el valor de R^2 estará siempre entre 0 y 1. Por esto, puede interpretarse como la **proporción de variabilidad de la respuesta que queda explicada** por la separación de las observaciones en grupos a través del modelo lineal.

Es importante observar también que el contraste ANOVA, como todos los contrastes de hipótesis estadísticos, posee una probabilidad o riesgo controlado de rechazar la hipótesis nula H_0 cuando esta es cierta, o dicho de manera más técnica, de cometer **errores de tipo I**. Como hemos visto, esta probabilidad de error tipo I es precisamente el nivel de significación α, que puede ser fijado por el analista. Nos podríamos preguntar entonces por qué no tomar siempre un α lo más bajo posible, digamos $\alpha = 0$, que nos protegería totalmente de cometer estos errores de tipo I. La respuesta es que la protección contra el riesgo de error tipo I tiene un precio, y es estar menos protegidos contra los **errores de tipo II**, que consisten en no rechazar H_0 cuando

esta es falsa. Nótese que los errores de tipo II están relacionados con la sensibilidad del contraste ANOVA para detectar diferencias entre las medias de la respuesta en los diferentes niveles del factor. Esta sensibilidad, que en lenguaje técnico se conoce como **potencia del contraste**, es directamente proporcional al nivel de significación α, de manera que un α más bajo implica a su vez una menor potencia del ANOVA a la hora de detectar una diferencia dada entre esas medias. Por ello, es necesario siempre lograr un equilibrio adecuado entre el riesgo de cometer ambos tipos de errores. No obstante, para un α fijo y unas diferencias reales entre las medias también fijas, es posible aumentar la potencia del ANOVA de dos maneras:

1) Aumentando n, esto es, el número de observaciones o repeticiones del experimento que se llevan a cabo en cada nivel del factor. Al replicar el experimento, las medias obtenidas son más precisas, esto es, tienen menor varianza, y esto conlleva una mayor sensibilidad del ANOVA para detectar unas diferencias dadas entre las medias. No obstante, esto puede no ser siempre posible o viable, ya que implica un mayor coste económico del experimento, y por otro lado aumentar la replicación puede también aumentar la heterogeneidad de la muestra, por lo que incluso si es viable económicamente este procedimiento puede no ser conveniente o eficaz.

2) Reduciendo σ^2, lo que implica disminuir el error experimental. La manera de conseguir esto suele pasar por controlar experimentalmente algunos de los factores que influyen en la respuesta, y cuyo efecto se acumula en el error experimental si no son controlados y planificados. Al controlarlos, se elimina su efecto de σ^2, y el experimento será más preciso y eficaz a la hora de detectar diferencias entre los niveles del factor de interés. Esto conlleva la utilización de modelos estadísticos con variables bloque o con más de un factor, que serán el objeto de las siguientes secciones de este tema.

1.2.6. Contrastes múltiples

Si el test ANOVA concluye que se ha de rechazar la hipótesis nula H_0, podemos tener cierta confianza en que el factor *X* tiene influencia sobre la respuesta *Y*, y que algunos niveles del factor proporcionarán en promedio valores de la respuesta diferentes a los de otros niveles. Sin embargo, el ANOVA no nos informa de qué niveles o grupos son diferentes entre sí, ya que solamente contrasta si existe al menos un grupo significativamente diferente al resto, sin entrar en la cuestión de cuáles son los grupos efectivamente diferentes. De hecho, no existe un procedimiento que informe directamente de todas las diferencias entre grupos, y la única posibilidad es analizar las diferencias entre grupos dos a dos, esto es, entre pares de grupos.

La comparación de las medias de dos grupos de observaciones es un procedimiento estándar y bien conocido, típicamente realizado mediante el denominado contraste de hipótesis de la *t* de Student para la comparación de dos muestras. Así, dados dos grupos *i* y *j*, *i*,*j* = 1,…,*a*, con sus correspondientes muestras $(y_{i1},...,y_{in})$ y $(y_{j1},...,y_{jn})$, en este contraste se parte de la hipótesis nula $H_0^{i,j}: \mu_i = \mu_j$, bajo la cual se puede comprobar que el estadístico

$$t_0^{i,j} = \frac{\overline{y}_{i\cdot} - \overline{y}_{j\cdot}}{\sqrt{MCR\frac{2}{n}}} \sim t_{N-a}$$

y se rechazará la hipótesis nula cuando $|t_0^{i,j}| > t_{N-a;\alpha/2}$, con α el nivel de significación del test.

La aplicación reiterada de este contraste para distintos pares de medias presenta sin embargo un problema, conocido como el **problema de los contrastes múltiples**. Para ilustrarlo, supongamos que se tienen $a = 6$ grupos, y que se aplica el contraste *t* anterior a cada una de los

$\binom{6}{2}=15$ pares posibles de medias de esos grupos. Si se toma α = 0.05 como nivel de riesgo de error de tipo I, se tiene que, para cualquier par de grupos *i,j* = 1,…,6, la probabilidad de no rechazar $H_0^{i,j}$ cuando esta es cierta viene dada por $P(|t_0^{i,j}|\leq t_{N-a;\alpha/2})=1-\alpha=0.95$, pero conjuntamente, la probabilidad de no cometer error de tipo I en ninguna de las 15 comparaciones, si estas fueran independientes, es entonces

$$P(|t_0^{i,j}|\leq t_{N-a;\alpha/2},\forall\{i,j\})=(1-\alpha)^{15}=0.95^{15}=0.4633$$

y no 1 – α = 0.95.

La conclusión que se desprende de lo anterior es que, **al aplicar reiteradamente un test, es bastante probable que, aunque no existan diferencias entre los grupos, estas aparezcan en los contrastes como consecuencia de la aleatoriedad presente en cualquier experimento**. De hecho, si el número de comparaciones es muy grande, es esperable que al menos una comparación indique la diferencia entre dos grupos, aunque todos sean en realidad iguales. Esto es consecuencia de que, si hacemos muchas pruebas, un suceso poco probable terminará ocurriendo. Por ejemplo, aunque la probabilidad de obtener cinco caras seguidas al lanzar una moneda bien hecha es baja, es casi seguro que, si lanzamos la moneda un número suficientemente grande de veces, en algún momento observaremos cinco caras seguidas.

Se han estudiado y desarrollado una importante cantidad de procedimientos para resolver o aliviar este problema de los contrastes múltiples. Aquí veremos un **procedimiento basado en la desigualdad de Bonferroni**, que es a la vez simple y general, aunque no siempre óptimo.

Sea *c* el número total de comparaciones que se desea realizar. Si tenemos *a* grupos y se quieren realizar comparaciones entre cada par de grupos, será $c=\binom{a}{2}$. Para todo *k* = 1,…,*c*, consideremos el suceso A_k = *rechazar* $H_0^{i,j}$ *cuando es cierta*. Supongamos que las comparaciones entre medias se hacen con un nivel de significación constante α, de manera que $P(A_k)=\alpha\ \forall k=1,...,c$. Sea ahora *B* el suceso *rechazar una o más de las hipótesis* $H_0^{i,j}$ *cuando todas las medias son iguales*. Entonces, *B* es la unión de los sucesos A_k, esto es, $B=A_1\cup...\cup A_c$. Nótese que los sucesos A_k no son mutuamente excluyentes, esto es, $A_k\cap A_{k'}\neq\varnothing$, por lo que se puede aplicar la desigualdad de Bonferroni

$$P(B)=P(A_1\cup...\cup A_c)\leq\sum_{k=1}^{c}P(A_k)=c\alpha$$

Así pues, si se pretende garantizar una probabilidad de error tipo I total α_T para el conjunto de *c* comparaciones, la probabilidad del suceso *B* debe ser como máximo α_T, lo que puede conseguirse imponiendo en cada contraste individual un nivel de significación α tal que

$$\alpha=\frac{\alpha_T}{c}$$

Este procedimiento se conoce de hecho como el **método de Bonferroni** para comparaciones múltiples. A pesar de su simplicidad puede ser muy útil en la práctica y es aplicable en un rango muy general de circunstancias. Sin embargo, hay que tener en cuenta que se trata de un procedimiento aproximado, en el sentido de que la cota superior obtenida para *P*(*B*) puede estar lejos de ser la óptima. Otros muchos métodos intentan reducir esta cota, aunque en muchos

casos a costa de introducir algunos supuestos que limitan su generalidad, y por tanto su aplicación.

1.2.7. Diagnosis del modelo

Una vez construido el modelo, contrastada la hipótesis de igualdad de medias mediante el test ANOVA y estimadas las diferencias entre las medias de los grupos, conviene estudiar si los supuestos básicos introducidos en la Sección 1.2.1 son razonables en el problema concreto bajo estudio. Para ello, es necesario obtener los residuos $e_{ij} = y_{ij} - \overline{y}_{i\cdot}$, que son las estimaciones de las perturbaciones o errores u_{ij}. De acuerdo con los supuestos básicos, los errores u_{ij} deben ser valores aleatorios de una $N(0,\sigma^2)$ e independientes.

Conviene recordar que, al haber estimado a parámetros, únicamente $N - a$ de los N residuos e_{ij} son independientes, ya que su suma en cada grupo i debe ser cero. Sin embargo, si N es grande con relación a a, es posible tratarlos como aproximadamente independientes y comprobar empíricamente, casi siempre gráficamente, su comportamiento. En especial, es siempre conveniente comprobar:

a) Si la distribución de los e_{ij} es normal, lo que puede hacerse mediante gráficos de probabilidad normal y contrastes de normalidad.

b) Si existen valores atípicos u *outliers*, esto es, residuos e_{ij} anormalmente grandes o pequeños con relación a los demás e_{ij}. Cuando esto ocurre debe buscarse la causa de esta anormalidad y, en caso de duda, desechar la observación de la que proviene si se sospecha que corresponde con un error de datos o a cambios imprevistos en las condiciones experimentales. Un diagrama de puntos o un histograma son herramientas adecuadas para estudiar visualmente la distribución de los residuos.

c) Si la variabilidad es constante en todos los grupos, esto es, si se verifica el supuesto de homocedasticidad. Para ello es conveniente graficar los residuos e_{ij} frente a las medias $\overline{y}_{i\cdot}$ de cada grupo (los valores predichos), para comprobar que su variabilidad no depende del nivel medio de la respuesta $\overline{y}_{i\cdot}$ y se mantiene relativamente estable en todos los grupos.

d) La independencia puede estudiarse mediante un gráfico de los e_{ij} frente al orden de obtención de las observaciones, que permite detectar posibles tendencias u otros patrones no aleatorios inesperados.

¿Qué ocurre si tras algunas de estas comprobaciones pensamos que alguno de los supuestos básicos podría no estarse cumpliendo? La respuesta es que depende. La falta de normalidad en los errores tiene en general poca influencia en el contraste ANOVA y en las comparaciones entre las medias, ya que estas tendrán siempre una distribución próxima a la normal por el teorema central del límite. Por tanto, los resultados de estos contrastes serán sustancialmente válidos aunque los datos sean no normales, y en este sentido es posible afirmar que **el análisis de la varianza es una técnica *robusta* frente a desviaciones de la normalidad**. No obstante, cuando la no normalidad se produce por una mayor acumulación de probabilidad en las colas de la distribución de los errores, esta desviación sí puede tener consecuencias graves, ya que los estimadores de mínimos cuadrados son especialmente sensibles e inestables en este contexto. En este caso, la mejor solución es **transformar los datos**, como veremos un poco más abajo, de modo que la distribución de las observaciones transformadas sea más próxima a la normal.

El efecto de la no verificación de la homocedasticidad de los errores puede ser importante, especialmente si las diferencias entre las varianzas de los grupos son muy pronunciadas (una

relación de 1 a 5 o mayor) o hay diferencias marcadas en el número de observaciones realizadas en distintos grupos. En ambos casos, suele ser aconsejable también llevar a cabo una transformación de los datos para estabilizar la varianza y aproximarse a la homocedasticidad, aunque en el segundo caso con tamaños muestrales diferentes en cada grupo esto no arregla totalmente el problema. Por esto, es altamente recomendable que el diseño del experimento sea **balanceado**, esto es, que se recoja el mismo número *n* de observaciones en cada uno de los grupos.

Finalmente, **la falta de independencia entre las observaciones puede tener un efecto muy grave sobre las conclusiones del ANOVA**, ya que las fórmulas de las varianzas de las medias muestrales $\overline{y}_{i\cdot}$ de los grupos se invalidan en este caso, y por tanto todos las estimaciones de la precisión de los estimadores serán erróneas. La manera más eficaz de prevenir la dependencia de las observaciones es la **aleatorización del experimento**, lo que incide de nuevo en la necesidad de un cuidadoso diseño del experimento para permitir un posterior análisis adecuado de los datos recogidos.

1.2.8. Transformaciones para conseguir homocedasticidad

El modelo básico estudiado se basa como sabemos en el supuesto de homocedasticidad, esto es, de que las observaciones en los distintos grupos siguen una misma distribución que solo difiere de un grupo a otro por el valor de su media. Sin embargo, en la práctica no es infrecuente que los grupos difieran no solamente en la media, sino también en su variabilidad. Por ejemplo, supongamos que las observaciones se generan realmente mediante el modelo

$$y_{ij} = \mu_i u_{ij} \;\; i = 1,...,a, j = 1,...,n$$

en el que a diferencia del modelo (1.1) las perturbaciones u_{ij} son variables aleatorias con media 1 y varianza constante que poseen un efecto sobre las observaciones de carácter multiplicativo en lugar de aditivo. Este modelo producirá heterocedasticidad, ya que los grupos con media μ_i más alta tendrán una mayor variabilidad. Sin embargo, realizando una transformación logarítmica de las observaciones, esto es, $z_{ij} = \log(y_{ij}) = \log(\mu_i) + \log(u_{ij}) \;\forall i, j$, se obtiene un modelo aditivo como (1.1) en el que las nuevas perturbaciones $v_{ij} = \log(u_{ij})$ tendrán media 0 y seguirán teniendo varianza constante, y por tanto las observaciones transformadas z_{ij} serán homocedásticas. Además, la heterocedasticidad del modelo multiplicativo anterior suele ir unida a una distribución no normal asimétrica de los errores u_{ij}, por lo que la transformación logarítmica puede arreglar simultáneamente ambos problemas.

Para detectar si es necesario transformar las observaciones son útiles los siguientes pasos.

a) Estudiar gráficamente las distribuciones de los residuos e_{ij}. Si su distribución es muy asimétrica, convendrá transformar las observaciones para convertirlas en normales.
b) Construir el diagrama de dispersión de los residuos e_{ij} frente a las medias $\overline{y}_{i\cdot}$ de cada grupo. Si se dispone de unas n = 5 o más observaciones por grupo, suele ser más clarificador calcular las desviaciones típicas muestrales s_i de cada grupo y graficarlas en el eje *Y* frente a las medias $\overline{y}_{i\cdot}$ en el eje X. Si se observa una relación entre ambas, será conveniente transformar para lograr homocedasticidad.

¿Qué transformación hay que aplicar? Lo mejor para responder esta pregunta es observar el tipo de relación entre las medias y la variabilidad, y centrarse en encontrar una transformación que elimine esa dependencia. Puede demostrarse que si transformamos una variable aleatoria *y* como *z* = *h*(*y*), se verifica que las desviaciones típicas de la variable original y la transformada

están aproximadamente relacionadas mediante la expresión $\sigma_z \cong \sigma_y |h'(y)|$. Por tanto, si queremos que la desviación típica de la variable transformada *z* sea una constante *k*, debe verificarse que $|h'(y)| = k / \sigma_y$.

Por ejemplo, supongamos que la relación observada entre las medias $\overline{y}_{i\cdot}$ y las desviaciones típicas s_i es del tipo $s_i = \overline{y}_{i\cdot}^{\,p}$, con $p \in \mathbb{R}$. Entonces, podemos estimar la transformación *h* que convierte la variable *y* en otra *z* con varianza constante como

$$h(y) = \int \frac{k}{y^p} dy = k' y^{1-p}$$

Así, si la variabilidad está ligada a la media con una relación del tipo $s_i = \overline{y}_{i\cdot}^{\,p}$, transformando las observaciones con $h(y) = y^{1-p}$ se obtendrán nuevas observaciones con varianza aproximadamente constante.

Este razonamiento conduce a considerar de manera general transformaciones de tipo potencial dependientes de un parámetro λ, que constituyen la familia de **transformaciones Box-Cox**:

$$h(y) = y^{(\lambda)} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln y & \lambda = 0 \end{cases}$$

El valor apropiado de λ a usar se puede estimar calculando la variable transformada $y^{(\lambda)}$ para diversos valores de λ y obteniendo los residuos $e_{ij}^{(\lambda)}$ correspondientes tras el ajuste del modelo ANOVA. El valor de λ que proporcione una SCR menor será el valor a emplear en la transformación. Para relaciones del tipo $s_i = \overline{y}_{i\cdot}^{\,p}$ se ha de tomar $\lambda = 1 - p$.

1.3. DISEÑO UNIFACTORIAL CON BLOQUES

Vamos ahora a estudiar un modelo de diseño experimental algo más complejo que el caso unifactorial simple, en el que además de un factor de interés se introduce un segundo factor, sobre el que en principio no se tiene interés en demostrar su influencia sobre la variable respuesta, pero se quiere controlar su efecto de manera que no distorsione el efecto del factor de interés. Como vimos en la Sección 1.1, este tipo de factores empleados para controlar la variabilidad de la respuesta sin que tengan interés por si mismos recibían el nombre de **variables bloque**. Así pues, en esta sección se estudiará el caso de un diseño experimental en que se considera una variable bloque junto a un único factor de interés.

1.3.1. Modelo unifactorial con bloques aleatorizados

Retomando el ejemplo introducido en la Sección 1.1 sobre el estudio del efecto de un conjunto de métodos de forjado sobre la dureza de una aleación en un proceso metalúrgico, supongamos que, conociendo que el operador que realiza el proceso puede influir en la dureza medida, se decide controlar estadísticamente su efecto para conseguir comparaciones entre los métodos de forjado independientes del efecto del operador. De esta manera, esperamos que las posibles diferencias entre los métodos de forjado sean más claramente visibles. Por tanto, en este caso estaríamos considerando un factor de interés, que como antes viene dado por la variable *método de forjado*, así como una variable bloque, dada por el *operador*.

Supongamos entonces que se tienen *a* métodos de forjado y que estos son llevados a cabo por *b* operadores diferentes, y que se observa la respuesta (dureza del espécimen producido) una única vez para cada combinación método-operador, esto es, para el **cruce** de los niveles del

factor con los niveles del bloque. De esta forma, los datos vendrán dispuestos en forma de tabla de doble entrada, como se ilustra en la Figura 1.3, donde y_{ij} denota el valor de la respuesta observado en la combinación del nivel *i* del factor (método de forjado) con el nivel *j* del bloque (operador). Llamaremos **tratamiento** a cada una de estas combinaciones de un nivel del factor con un nivel del bloque, y por tanto el experimento consistirá en observar la respuesta en cada uno de los *ab* tratamientos.

		Operador			
		1	2	…	*b*
Método	1	y_{11}	y_{12}	…	y_{1b}
	2	y_{21}	y_{22}	…	y_{2b}
	⋮	⋮	⋮	..	⋮
	a	y_{a1}	y_{a2}	…	y_{ab}

Figura 1.3: Formato de los datos en un experimento unifactorial aleatorizado por bloques

Aplicando el principio de aleatorización, se supondrá que el orden en que cada operador (i.e., nivel de la variable bloque) aplica cada uno de los métodos de forjado (i.e. nivel del factor) se determina aleatoriamente, para evitar sesgos derivados del aprendizaje de los operadores, de cambios en las condiciones del horno, etc. Esto es, la aleatorización de los niveles del factor en este caso se realiza dentro de cada nivel de la variable bloque. Esto lleva a denominar a este tipo de diseño experimental como **diseño unifactorial aleatorizado por bloques**, en contraste con el diseño unifactorial simple o completamente aleatorizado estudiado en la anterior Sección 1.2. Subrayemos de nuevo que en principio solo se realiza una observación de la respuesta para cada tratamiento o combinación método-operador, por lo que se tendrá un total de $N = ab$ observaciones. Además, se supondrá que el efecto combinado de un método y un operador depende aditivamente de los efectos individuales del método y del operador. De este modo, se llega al siguiente **modelo estadístico** de las observaciones de la respuesta:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad i = 1,...,a,\ j = 1,...,b \tag{1.3}$$

Así pues, este modelo descompone la respuesta observada como suma de las siguientes componentes:

- Un efecto global μ, que recoge el nivel promedio de la respuesta para todas las condiciones experimentales.
- Un efecto de los niveles del factor de interés α_i, que recoge el efecto incremental del nivel *i* del factor en relación al promedio general. Se supone por tanto que se cumple la restricción $\sum_{i=1}^{a} \alpha_i = 0$.
- Un efecto de los niveles de la variable bloque β_j, que recoge el efecto incremental del nivel *j* de la variable bloque. También se supone que $\sum_{j=1}^{b} \beta_j = 0$.
- Un efecto aleatorio ε_{ij}, que recoge el efecto de todas las restantes causas de variabilidad de la respuesta no consideradas en el experimento. Como en el modelo unifactorial simple descrito en la Sección 1.2, estos efectos aleatorios también reciben el nombre de errores experimentales, ruido o perturbaciones aleatorias, y se supone que se comportan como variables aleatorias normales independientes de media 0 y con la misma dispersión en cualquier combinación factor-bloque, esto es, $\varepsilon_{ij} \sim N(0,\sigma^2)\ \forall i, j$,

su distribución conjunta es normal multivariante y se tiene que $Cov[\varepsilon_{ij}, \varepsilon_{i',j'}] = 0$ si $(i, j) \neq (i', j')$.

Por tanto, este modelo (1.3) parte de los mismos supuestos básicos sobre los errores aleatorios ε_{ij} que los modelos unifactoriales simples (1.1) y (1.2) tratados en la Sección 1.2. Esto es, estos errores se suponen normales independientes con media 0 y varianza constante σ^2 . Por tanto, las observaciones y_{ij} condicionadas a cada tratamiento también conservarán su distribución normal con media $E[y_{ij}] = \mu + \alpha_i + \beta_j$ y varianza $V[y_{ij}] = \sigma^2$, y se mantiene la independencia entre observaciones diferentes.

Nótese también que el modelo anterior contiene a primera vista 1 + *a* + *b* parámetros: la media global μ, *a* efectos α_i y *b* efectos β_j . Sin embargo, por las restricciones consideradas, realmente hay *a* – 1 efectos independientes del factor y *b* – 1 efectos independientes del bloque. Además, la varianza constante σ^2 constituye también un parámetro. Por tanto, en total se tendrían 1 + (*a* – 1) + (*b* – 1) +1 = *a* + *b* parámetros independientes.

Como se ha indicado, el objetivo fundamental al analizar datos con esta forma es discernir si el factor de interés tiene influencia sobre la respuesta, lo que equivaldrá a contrastar la hipótesis nula $H_0(\alpha): \alpha_1 = ... = \alpha_a = 0$. También puede tener interés contrastar el efecto de la variable bloque, lo que vendrá asociado a la hipótesis nula $H_0(\beta): \beta_1 = ... = \beta_b = 0$.

1.3.2. Estimación de los parámetros del modelo

En base al modelo (1.3) es posible aplicar el método de mínimos cuadrados para obtener estimadores de los parámetros μ, α_i y β_j . Como en la Sección 1.2.2, este método procede mediante la minimización de la función de pérdida dada por la suma de los cuadrados de los errores $\varepsilon_{ij} = y_{ij} - (\mu + \alpha_i + \beta_j)$, esto es, minimizando la función

$$L(\mu, \alpha_1, ..., \alpha_a, \beta_1, ..., \beta_b) = \sum_{i=1}^{a}\sum_{j=1}^{b} \varepsilon_{ij}^2 = \sum_{i=1}^{a}\sum_{j=1}^{b} (y_{ij} - (\mu + \alpha_i + \beta_j))^2$$

Es directo comprobar que, planteando y resolviendo el sistema de ecuaciones normales con las restricciones impuestas sobre los parámetros al describir el modelo (1.3), se obtienen los siguientes estimadores de mínimos cuadrados (ejercicio):

$$\hat{\mu} = \overline{y}_{..}$$

$$\hat{\alpha}_i = \overline{y}_{i.} - \overline{y}_{..}, \ i = 1, ..., a$$

$$\hat{\beta}_j = \overline{y}_{.j} - \overline{y}_{..}, \ j = 1, ..., b$$

donde ahora se toma la siguiente notación:

$$\overline{y}_{..} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij} \ ; \ \overline{y}_{i.} = \frac{1}{b}\sum_{j=1}^{b} y_{ij} \ ; \ \overline{y}_{.j} = \frac{1}{a}\sum_{i=1}^{a} y_{ij}$$

Estos estimadores cumplen propiedades muy similares a las demostradas en la Sección 1.2.2 para los estimadores de mínimos cuadrados del modelo unifactorial simple.

Proposición 1.4. Los estimadores de mínimos cuadrados anteriores tienen las siguientes distribuciones:

1) $\hat{\mu} = \overline{y}_{..} \sim N(\mu, \frac{\sigma^2}{ab})$

2) $\hat{\alpha}_i = \overline{y}_{i.} - \overline{y}_{..} \sim N(\alpha_i, \frac{a-1}{ab}\sigma^2)$

3) $\hat{\beta}_j = \overline{y}_{.j} - \overline{y}_{..} \sim N(\beta_j, \frac{b-1}{ab}\sigma^2)$

Demostración: Ejercicio.

A partir de las estimaciones anteriores de los parámetros es posible definir las nociones de **valores predichos** por el modelo lineal y sus **residuos**.

Definición 1.5. Se denominan ***valores predichos*** del modelo unifactorial con bloques aleatorizados al resultado de sustituir, para cada observación, los parámetros en la expresión (1.3) por los correspondientes estimadores de mínimos cuadrados. Esto es, el valor predicho por el modelo para la observación $y_{ij}, i = 1,...,a, j = 1,...,b$ viene dado por

$$\hat{y}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j = \overline{y}_{i.} + \overline{y}_{.j} - \overline{y}_{..}, \quad i = 1,...,a, j = 1,...,b$$

Definición 1.6. Se denominan ***residuos*** del modelo unifactorial con bloques aleatorizados a la diferencia entre los valores observados y los valores predichos por ese modelo, esto es,

$$e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - (\overline{y}_{i.} + \overline{y}_{.j} - \overline{y}_{..}),\ i = 1,...,a,\ j = 1,...,b$$

Como antes, podemos interpretar los residuos como estimaciones de los errores $\varepsilon_{ij} = y_{ij} - (\mu + \alpha_i + \beta_j)$. Es importante observar que los residuos no son independientes, ya que están sometidos a ciertas restricciones derivadas de la estimación de los parámetros. En particular, sumando todos los residuos asociados respectivamente a cada nivel *i* del factor y a cada nivel *j* del bloque, se obtiene

$$\sum_{j=1}^{b} e_{ij} = 0 \qquad i = 1,...,a$$

$$\sum_{i=1}^{a} e_{ij} = 0 \qquad j = 1,...,b$$

Por tanto, se tienen *a* + *b* restricciones para los *ab* residuos, aunque una de estas restricciones es dependiente de las *a* + *b* – 1 restantes (ejercicio), y por tanto existirán en total *ab* – (*a* + *b* – 1) = (*a* – 1) (*b* – 1) residuos no determinados. Luego en este caso los residuos poseerán (*a* – 1) (*b* – 1) **grados de libertad**.

1.3.3. Análisis de la varianza en diseños unifactoriales con bloques aleatorizados

Como se introdujo en la Sección 1.3.1, el objetivo que se persigue con un experimento unifactorial con una variable bloque es determinar si el factor considerado tiene influencia sobre la respuesta, lo que se traduce en contrastar la hipótesis nula $H_0(\alpha): \alpha_1 = ... = \alpha_a = 0$. Como en el caso del modelo unifactorial simple visto en la Sección 1.2, para contrastar esta hipótesis es necesario descomponer primero la variabilidad de la respuesta en las diversas fuentes de variación. En el caso del modelo unifactorial con bloques, sin embargo, el diseño del experimento permite separar la variabilidad total en tres fuentes: la debida al factor, la debida a la variable bloque y la residual. Una vez obtenida esta descomposición, siguiendo la metodología ANOVA expuesta en la Sección 1.2.5 el contraste de H_0 se realiza mediante una

comparación entre la variabilidad de la respuesta explicada por el factor y la variabilidad residual o no explicada.

Igualmente, de cara a valorar la influencia de la variable bloque sobre la respuesta se ha de contrastar la hipótesis nula $H_0(\beta): \beta_1 = ... = \beta_b = 0$, lo que pasa por comparar la variabilidad de la respuesta que explica la variable bloque con la misma variabilidad residual que en el caso de la comparación anterior para el efecto del factor.

La descomposición de la variabilidad total a partir del modelo unifactorial con bloques parte de la siguiente igualdad, válida para cada una de las *N* = *ab* observaciones:

$$\frac{1}{\sigma}(y_{ij} - \mu) = \frac{1}{\sigma}(\overline{y}_{..} - \mu) + \frac{1}{\sigma}(\overline{y}_{i.} - \overline{y}_{..}) + \frac{1}{\sigma}(\overline{y}_{.j} - \overline{y}_{..}) + \frac{1}{\sigma}(y_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}_{..})$$

Podemos expresar esta identidad simultáneamente para las *N* observaciones de manera vectorial como

$$\mathbf{Y} = \mathbf{S}_1 + \mathbf{S}_2 + \mathbf{S}_3 + \mathbf{S}_4 \tag{1.4}$$

donde todos los vectores en negrita tienen longitud *N* y vienen dados por

$$\mathbf{Y}^t = \frac{1}{\sigma}(y_{11} - \mu, ..., y_{ab} - \mu)$$

$$\mathbf{S}_1^t = \frac{1}{\sigma}(\overline{y}_{..} - \mu, ..., \overline{y}_{..} - \mu)$$

$$\mathbf{S}_2^t = \frac{1}{\sigma}(\overline{y}_{1.} - \overline{y}_{..}, ..., \overline{y}_{1.} - \overline{y}_{..}, \overline{y}_{2.} - \overline{y}_{..}, ..., \overline{y}_{2.} - \overline{y}_{..}, ..., \overline{y}_{a.} - \overline{y}_{..}, ..., \overline{y}_{a.} - \overline{y}_{..})$$

$$\mathbf{S}_3^t = \frac{1}{\sigma}(\overline{y}_{.1} - \overline{y}_{..}, \overline{y}_{.2} - \overline{y}_{..}, ..., \overline{y}_{.b} - \overline{y}_{..}, \overline{y}_{.1} - \overline{y}_{..}, \overline{y}_{.2} - \overline{y}_{..}, ..., \overline{y}_{.b} - \overline{y}_{..}, ..., \overline{y}_{.1} - \overline{y}_{..}, \overline{y}_{.2} - \overline{y}_{..}, ..., \overline{y}_{.b} - \overline{y}_{..})$$

$$\mathbf{S}_4^t = \frac{1}{\sigma}(e_{11}, ..., e_{ab})$$

En particular, nótese que el vector **Y** tiene *N* grados de libertad, en tanto las observaciones son independientes. Los *N* elementos del vector $\mathbf{S}_1$ son iguales a $\overline{y}_{..} - \mu$, por lo que este vector tiene solo 1 grado de libertad. El vector $\mathbf{S}_2$ contiene *a* valores distintos $\hat{\alpha}_i = \overline{y}_{i.} - \overline{y}_{..}$, *i* = 1,…,*a*, cada uno repetido *b* veces, aunque al cumplirse que $\sum_{i=1}^{a} \hat{\alpha}_i = 0$ (ejercicio) solo tendrá *a* – 1 grados de libertad. Del mismo modo, el vector $\mathbf{S}_3$ contiene *b* valores distintos $\hat{\beta}_j = \overline{y}_{.j} - \overline{y}_{..}$, *j* = 1,…,*b*, cada uno repetido *a* veces, aunque al cumplirse la restricción $\sum_{j=1}^{b} \hat{\beta}_j = 0$ (ejercicio) solo tendrá *b* – 1 grados de libertad. Y finalmente, el vector de residuos $\mathbf{S}_4$ contiene los *N* residuos estimados $e_{ij} = y_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}_{..}$, que deben sumar 0 por filas y columnas y tiene por tanto *N* – (*a* + *b* – 1) grados de libertad.

No es complicado comprobar que los vectores $\mathbf{S}_1$,…,$\mathbf{S}_4$ son ortogonales dos a dos puesto que los correspondientes productos escalares se anulan (ejercicio):

$$\mathbf{S}_1^t \mathbf{S}_2 = 0$$

$$\mathbf{S}_1^t \mathbf{S}_3 = 0$$

$$\mathbf{S}_1^t\mathbf{S}_4 = 0$$

$$\mathbf{S}_2^t\mathbf{S}_3 = 0$$

$$\mathbf{S}_2^t\mathbf{S}_4 = 0$$

$$\mathbf{S}_3^t\mathbf{S}_4 = 0$$

En base a esta descomposición ortogonal, y en tanto que (1.4) puede expresarse como $\mathbf{Y} - \mathbf{S}_1 = \mathbf{S}_2 + \mathbf{S}_3 + \mathbf{S}_4$, tomando cuadrados de los módulos a ambos lados de esta igualdad se obtiene la identidad

$$|\mathbf{Y} - \mathbf{S}_1|^2 = |\mathbf{S}_2|^2 + |\mathbf{S}_3|^2 + |\mathbf{S}_4|^2$$

que se puede expresar de manera equivalente como

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

o más simplificadamente

$$\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \overline{y}_{..})^2 = b\sum_{i=1}^{a}\hat{\alpha}_i^2 + a\sum_{j=1}^{b}\hat{\beta}_j^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}e_{ij}^2$$

Esta expresión es la **descomposición en suma de cuadrados** asociada al modelo unifactorial con una variable bloque, en la que los diferentes términos reciben la siguiente interpretación:

- $SCT = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \overline{y}_{..})^2$ se denomina **suma de cuadrados total**, como en el modelo unifactorial simple, y representa la variabilidad total observada de la respuesta alrededor de su media.
- $SCF = b\sum_{i=1}^{a}\hat{\alpha}_i^2$ es la **suma de cuadrados del factor**, y representa la parte de variabilidad de la respuesta explicada por las variaciones controladas en los niveles del factor.
- $SCB = a\sum_{j=1}^{b}\hat{\beta}_j^2$ es la **suma de cuadrados del bloque**, y representa la parte de variabilidad de la respuesta explicada por las variaciones controladas en los niveles de la variable bloque.
- $SCR = \sum_{i=1}^{a}\sum_{j=1}^{b}e_{ij}^2$ es la **suma de cuadrados de los residuos**, y representa la variabilidad residual o variabilidad de la respuesta no explicada por el modelo que conforma el error experimental.

Con esta terminología, la descomposición anterior puede expresarse sintéticamente como

$$SCT = SCF + SCB + SCR$$

Una vez establecida esta descomposición de la variabilidad total de la respuesta en las tres fuentes de variabilidad consideradas en el diseño unifactorial con una variable bloque, estamos en condiciones de formular el contraste ANOVA para este diseño experimental. Como en la Sección 1.2.5, la clave pasa por la aplicación del Teorema de Cochran sobre la identidad (1.4) anterior.

Para ello, es preciso ver que bajo la hipótesis de que ni el factor ni la variable bloque tienen influencia sobre la respuesta, esto es, bajo la hipótesis nula $H_0(\alpha,\beta): \alpha_i = 0\ \forall i;\ \beta_j = 0\ \forall j$, todas las observaciones y_{ij} tienen media μ y varianza σ^2. En otras palabras, bajo $H_0(\alpha,\beta)$ el vector **Y** en (1.4) se distribuye como $\mathbf{Y} \sim N(\mathbf{0},\mathbf{I})$, y se está por tanto en las condiciones del Teorema de Cochran, que garantiza entonces que las sumas de cuadrados anteriores son variables aleatorias independientes que se distribuyen como sigue:

$$\frac{\text{SCF}}{\sigma^2} \sim \chi^2_{a-1}, \quad \frac{\text{SCB}}{\sigma^2} \sim \chi^2_{b-1} \quad \text{y} \quad \frac{\text{SCR}}{\sigma^2} \sim \chi^2_{(a-1)(b-1)}$$

Cuando solo se asume que la variable bloque no tiene influencia sobre la respuesta, esto es, bajo $H_0(\beta): \beta_1 = ... = \beta_b = 0$, la distribución de SCB/σ^2 sigue siendo una chi-cuadrado independientemente de que el factor tenga influencia o no, esto es, de que los α_i sean o no nulos. En este caso la descomposición de la variabilidad total es análoga a la efectuada en la Sección 1.2.3, pero donde ahora la variabilidad residual se ha descompuesto en dos términos independientes, SCB y SCR.

Del mismo modo, si solo se asume que el factor no tiene influencia en la respuesta, es decir, bajo $H_0(\alpha): \alpha_1 = ... = \alpha_a = 0$, la distribución de SCF/σ^2 también sigue siendo una chi-cuadrado independientemente de si la variable bloque influye en la respuesta o no, esto es, de que los β_i sean o no nulos.

Por otro lado, es posible demostrar que la distribución de SCR/σ^2 siempre es chi-cuadrado, con independencia de que el factor o la variable bloque influyan o no en la respuesta.

En base a estos resultados es entonces posible plantear contrastes ANOVA similares a los realizados en la Sección 1.2.5 para dilucidar la cuestión de la influencia sobre la respuesta del factor y de la variable bloque. El contraste de mayor interés es el realizado sobre el factor, que parte de la hipótesis $H_0(\alpha): \alpha_1 = ... = \alpha_a = 0$, y se lleva a cabo mediante el estadístico de contraste

$$F_0(\alpha) = \frac{SCF/(a-1)}{SCR/[(a-1)(b-1)]} = \frac{MCF}{MCR} \overset{H_0(\alpha)}{\sim} F_{a-1,(a-1)(b-1)}$$

En consecuencia, se rechaza $H_0(\alpha)$ cuando $F_0 > F_{a-1,(a-1)(b-1),1-\alpha}$.

De manera similar, se puede contrastar la hipótesis $H_0(\beta): \beta_1 = ... = \beta_b = 0$ de que la variable bloque no tiene influencia mediante el estadístico de contraste

$$F_0(\beta) = \frac{SCB/(b-1)}{SCR/[(a-1)(b-1)]} = \frac{MCB}{MCR} \overset{H_0(\beta)}{\sim} F_{b-1,(a-1)(b-1)}$$

de manera que se rechaza $H_0(\beta)$ cuando $F_0 > F_{b-1,(a-1)(b-1),1-\alpha}$.

En la práctica, los paquetes de software estadístico suelen presentar la información relativa a este contraste ANOVA en un formato de tabla estandarizado, que se conoce como **tabla ANOVA del modelo unifacorial con bloques**, y que se muestra en la Figura 1.4.

FUENTE DE VARIABILIDAD	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	CUADRADOS MEDIOS	F_0	p-valor
FACTOR	SCF	$a-1$	MCF	MCF/MCR	$P(F_{a-1,(a-1)(b-1)} > F_0)$
BLOQUE	SCB	$b-1$	MCB	MCB/MCR	$P(F_{b-1,(a-1)(b-1)} > F_0)$
ERROR	SCR	$(a-1)(b-1)$	MCR		
TOTAL	SCT	$ab-1$			

Figura 1.4: Tabla ANOVA de un diseño unifactorial aleatorizado por bloques.

Un último apunte: es importante entender la diferencia en eficacia entre este diseño unifactorial con una variable bloque y el anterior diseño unifactorial simple. Si la influencia de la variable bloque es reducida o inexistente, el contraste $F_0(\alpha)$ del diseño unifactorial con bloque sería menos eficaz que el contraste ANOVA simple de la Sección 1.2.5, ya que el denominador en $F_0(\alpha)$ tendría menos grados de libertad, $(a-1)(b-1)$, que en el caso del contraste F del modelo unifactorial simple, que tiene $a(n-1)$ grados de libertad en el denominador, y donde se tendría que $n = b$. No obstante, siempre podemos transformar el diseño unifactorial con bloque en el diseño unifactorial simple, descartando simplemente la división por bloques, esto es, entendiendo que las b observaciones realizadas para cada uno de los a niveles del factor son simplemente replicaciones en lugar de observaciones bajo condiciones diferentes de la variable bloque.

Sin embargo, cuando la variable bloque tiene una influencia relevante sobre la respuesta, el diseño unifactorial con bloques es enormemente más eficaz que el diseño unifactorial simple. Si la SCB es grande, el término de variabilidad no explicada SCR será mucho menor, y el contraste $F_0(\alpha)$ será mucho más sensible a la hora de detectar la influencia del factor sobre la respuesta.

En resumen, esta comparación de eficacia entre ambos diseños muestra la conveniencia de cruzar siempre todos los niveles de las variables o factores con posible influencia sobre la respuesta: no se pierde precisión si algunos factores no tienen influencia, ya que podemos transformar el diseño cruzado en uno más simple, pero se puede ganar mucha precisión si realmente alguno de los factores o bloques cruzados sí tiene influencia.

1.4. DISEÑO BIFACTORIAL

Con frecuencia, un experimento tendrá como objetivo estudiar la influencia simultánea de varios factores de interés sobre una variable respuesta. Esto lleva naturalmente a la noción de **diseños multifactoriales**, en que se considera la variación planificada de más de un factor y se estudia no solo el efecto separado de cada uno de ellos sino también su posible **interacción**.

En esta sección estudiaremos el caso particular de un diseño con dos factores, que se denomina **diseño bifactorial**. Este tipo de diseño bifactorial es a primera vista similar a un diseño unifactorial con una variable bloque, en tanto se considera el cruce de los niveles de dos variables categóricas y la realización de observaciones aleatorizadas en cada combinación de niveles o tratamiento. Sin embargo, en términos del análisis estadístico subsiguiente, el modelo bifactorial introduce una diferencia importante, que es la consideración de un efecto de interacción entre ambos factores. Para poder tener en cuenta esta interacción, el modelo estadístico ha de incorporar un parámetro de interacción para cada tratamiento, lo que conlleva que el modelo resultante contenga más parámetros que tratamientos. Por ello, en el caso del modelo bifactorial, a diferencia del modelo unifactorial con bloque, siempre que se precise estudiar la posible interacción entre dos factores será necesario llevar a cabo la replicación del experimento, de modo que se obtengan al menos dos observaciones en cada tratamiento.

¿Qué significa que exista interacción entre dos factores? Básicamente, **existe interacción cuando el efecto de uno de los factores depende del nivel en que se encuentre el otro factor**. Por ejemplo, supongamos que se estudia el rendimiento de una reacción química bajo dos niveles (bajo y alto) de temperatura (*T*) y de concentración (*C*) de un reactivo. Denotemos estos niveles como *T*-, *T*+, *C*-, *C*+, y supongamos que en el tratamiento dado por la combinación de niveles (*T*-,*C*-) se observa un rendimiento de 10 unidades. Tomando este valor como base, si en el tratamiento (*T*-,*C*+) se observa un rendimiento de 15 unidades, se estimaría el efecto de pasar de un nivel bajo de concentración a uno alto en 15 – 10 = 5 unidades. Del mismo modo, si en el tratamiento (*T*+,*C*-) se observa un rendimiento de 13 unidades, estimaríamos el efecto de pasar de un nivel bajo de temperatura a uno alto en 13 – 10 = 3 unidades. Si el efecto de ambos factores fuese simplemente aditivo, en el tratamiento (*T*+,*C*+) se obtendría un rendimiento que combinaría los efectos de pasar de un nivel bajo a uno alto en temperatura y concentración, esto es, un rendimiento de 10 + 5 + 3 = 18 unidades. Gráficamente, podríamos resumir este comportamiento en el siguiente gráfico.

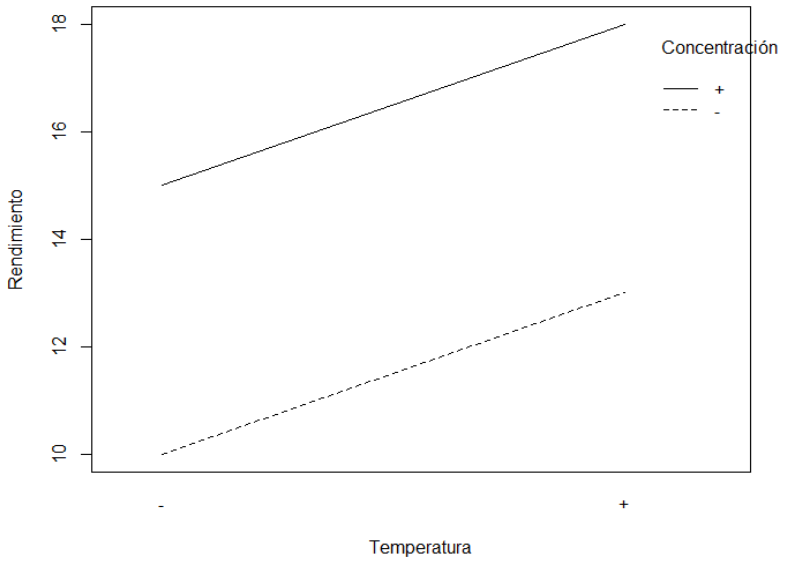

Figura 1.5: Gráfico de interacción en el que no se aprecia efecto de interacción.

Como se puede observar, el efecto de cambiar la temperatura de un nivel bajo a uno alto es el mismo en ambos niveles de concentración: en ambos casos, el rendimiento aumenta en 3 unidades. Y similarmente, el efecto de cambiar el nivel de concentración es aumentar el rendimiento en 5 unidades. Esto produce que en el gráfico anterior las dos líneas que muestran el efecto de la temperatura en ambos niveles de concentración sean paralelas. En esta situación, no existiría interacción entre los dos factores. Supongamos ahora que, en lugar de observar un rendimiento de 18 unidades, en el tratamiento (*T*+,*C*+) se observara un rendimiento de 7 unidades. En este caso, el gráfico de interacción resultante sería el que se muestra en la Figura 1.6 situada más abajo.

Nótese que ahora el efecto de pasar de un nivel bajo a uno alto de temperatura es de +3 unidades cuando la concentración es baja, y de -8 unidades cuando la concentración es alta. Del mismo modo, el efecto de aumentar el nivel de concentración sería de +5 cuando la temperatura es baja y de -6 cuando la temperatura es alta. En este caso, existiría una clara interacción entre ambos factores, lo que se traduce en las líneas no paralelas y que se cortan en el gráfico anterior.

Cuando existe interacción entre dos factores, utilizar un modelo que no la considere puede llevar a importantes errores en la estimación del efecto de los factores. Por ello, en un contexto de diseño multifactorial siempre es aconsejable introducir parámetros extra que puedan recoger

este efecto de interacción, y luego contrastar su significatividad estadística mediante los test apropiados.

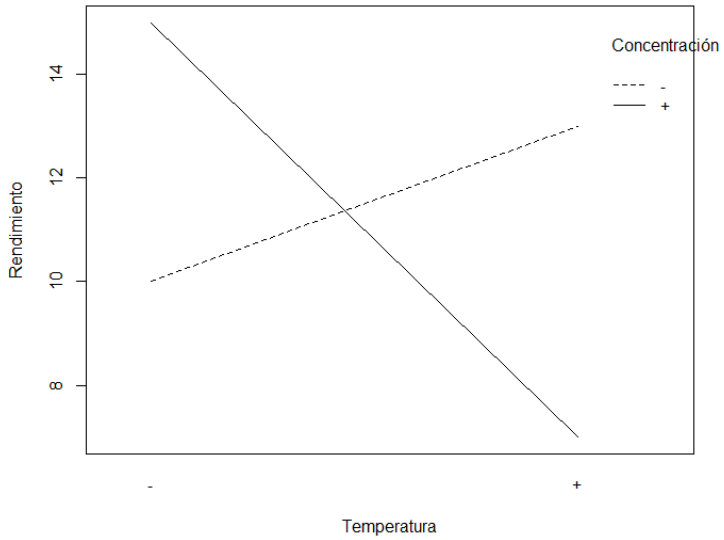

Figura 1.6: Gráfico de interacción mostrando una clara interacción entre los factores.

Así pues, supongamos que se quiere estudiar la influencia de dos factores sobre una variable respuesta, de manera que el primer factor tiene *a* niveles, el segundo tiene *b* niveles y en cada tratamiento o combinación de niveles de ambos factores se realizan *n* observaciones de la variable respuesta. Por tanto, en este diseño se tendría un total de $N = abn$ observaciones. La configuración típica de los datos en un diseño bifactorial se muestra en la Figura 1.7.

		Factor 2			
		1	2	…	*b*
Factor 1	1	$y_{111},\ldots,y_{11n}$	$y_{121},\ldots,y_{12n}$	…	$y_{1b1},\ldots,y_{1bn}$
	2	$y_{211},\ldots,y_{21n}$	$y_{221},\ldots,y_{22n}$	…	$y_{2b1},\ldots,y_{2bn}$
	⋮	⋮	⋮	..	⋮
	a	$y_{a11},\ldots,y_{a1n}$	$y_{a21},\ldots,y_{a2n}$	…	$y_{ab1},\ldots,y_{abn}$

Figura 1.7: Configuración del diseño bifactorial.

El **modelo estadístico** de las observaciones es entonces el siguiente:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \quad i = 1,\ldots,a,\ j = 1,\ldots,b,\ k = 1,\ldots,n$$

Así pues, este modelo descompone cada respuesta observada como suma de las siguientes componentes:

- Un efecto global μ, que recoge el nivel promedio de la respuesta para todas las condiciones experimentales.
- Un efecto α_i de los niveles del primer factor, que recoge el efecto incremental del nivel *i* de este factor en relación al promedio general.
- Un efecto β_j de los niveles del segundo factor, que recoge el efecto incremental del nivel *j* de este factor.
- Un término de interacción $(\alpha\beta)_{ij}$, que recoge el efecto de interacción entre ambos factores cuando el primero está en su nivel *i* y el segundo en el nivel *j*.

- Una perturbación aleatoria ε_{ijk}, que recoge el efecto de todas las restantes causas de variabilidad de la respuesta no consideradas en el experimento. Al igual que en los modelos anteriores, se supone que estas perturbaciones o errores se comportan como variables aleatorias normales independientes de media 0 y con la misma dispersión en cualquier tratamiento y replicación, esto es, $\varepsilon_{ijk} \sim N(0,\sigma^2)\ \forall i,j,k$, por lo que su distribución conjunta es normal multivariante y se tiene que $Cov[\varepsilon_{ijk},\varepsilon_{i'j'k'}]=0$ si $(i,j,k)\neq(i',j',k')$.

Las restricciones que se consideran sobre los anteriores parámetros son las siguientes:

$$\sum_{i=1}^{a}\alpha_i=\sum_{j=1}^{b}\beta_j=\sum_{i=1}^{a}(\alpha\beta)_{ij}=\sum_{j=1}^{b}(\alpha\beta)_{ij}=0$$

Esto indica que, como con los parámetros α y β asociados a los efectos aditivos de los factores, los parámetros de interacción $(\alpha\beta)$ miden también efectos incrementales respecto a la media global μ.

Por otro lado, para entender cómo se introduce la interacción en este tipo de modelo, tomemos esperanzas en la ecuación anterior del modelo:

$$E[y_{ijk}]=\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}\Leftrightarrow(\alpha\beta)_{ij}=E[y_{ijk}]-\mu-\alpha_i-\beta_j$$

Por tanto, el término de interacción se asimila a la diferencia entre el valor medio de la respuesta en el tratamiento (*i,j*) y el valor predicho suponiendo que el efecto de los dos factores solo fuese aditivo. En este sentido, la interacción mide la desviación de las observaciones de la aditividad de los efectos. En general, por tanto

Interacción = Respuesta observada – Respuesta prevista con aditividad

El número de parámetros independientes a estimar en este modelo es 1 (media μ) + (a – 1) parámetros α + $(b-1)$ parámetros β + $(a-1)(b-1)$ términos de interacción $(\alpha\beta)$ + 1 (varianza σ^2) = ab + 1 parámetros totales. Esto hace que la estimación sea imposible si no se realiza al menos un total de ab + 1 observaciones entre los ab tratamientos, aunque en la práctica el mínimo habitual es replicar al menos una vez cada observación de un tratamiento, lo que conduce a un mínimo de 2ab observaciones, que permite ya estimar todos los parámetros sin problemas.

La estimación de los parámetros anteriores se lleva a cabo mediante el método de mínimos cuadrados, minimizando la función

$$L(\mu,\alpha_1,\ldots,\alpha_a,\beta_1,\ldots,\beta_b,(\alpha\beta)_{11},\ldots,(\alpha\beta)_{ab})=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\varepsilon_{ijk}^2=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ij}-(\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}))^2$$

lo que conduce a las siguientes **estimadores de mínimos cuadrados**:

$$\hat{\mu}=\overline{y}_{...}$$

$$\hat{\alpha}_i=\overline{y}_{i..}-\overline{y}_{...},\ i=1,\ldots,a$$

$$\hat{\beta}_j=\overline{y}_{.j.}-\overline{y}_{...},\ j=1,\ldots,b$$

$$(\hat{\alpha\beta})_{ij}=\overline{y}_{ij.}-\overline{y}_{i..}-\overline{y}_{.j.}+\overline{y}_{...},\ \ i=1,\ldots,a,\ j=1,\ldots,b$$

La notación empleada para las medias es ahora la siguiente:

$$\overline{y}_{...} = \frac{1}{abn}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk} \;;\; \overline{y}_{i..} = \frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk} \;\;;\;\; \overline{y}_{.j.} = \frac{1}{an}\sum_{i=1}^{a}\sum_{k=1}^{n} y_{ijk} \;;\; \overline{y}_{ij.} = \frac{1}{n}\sum_{k=1}^{n} y_{ijk}$$

De este modo, para cualquier observación en el tratamiento (*i,j*), el **valor predicho** por el modelo en base a estas estimaciones será

$$\hat{y}_{ijk} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + (\alpha\hat{\beta})_{ij} = \overline{y}_{ij.}$$

Por tanto, los **residuos** del modelo vendrán dados por

$$e_{ijk} = y_{ijk} - \hat{y}_{ijk} = y_{ijk} - \overline{y}_{ij.}$$

y se tendrán $ab(n-1)$ residuos independientes, ya que la expresión anterior implica que los n residuos asociados a cada tratamiento (*i,j*) han de sumar 0.

En términos de descomposición de la variabilidad de la respuesta, es posible demostrar que en el modelo bifactorial se cumple la siguiente **descomposición en sumas de cuadrados**:

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\overline{y}_{...})^2}_{SCT} = \underbrace{bn\sum_{i=1}^{a}(\overline{y}_{i..}-\overline{y}_{...})^2}_{SCA} + \underbrace{an\sum_{j=1}^{b}(\overline{y}_{.j.}-\overline{y}_{...})^2}_{SCB} + \underbrace{n\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{y}_{ij.}-\overline{y}_{i..}-\overline{y}_{.j.}+\overline{y}_{...})^2}_{SCAB} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\overline{y}_{ij.})^2}_{SCR}$$

La demostración de esta descomposición procede comprobando que los diferentes términos en el término de la derecha de la igualdad

$$y_{ijk} - \overline{y}_{...} = (\overline{y}_{i..} - \overline{y}_{...}) + (\overline{y}_{.j.} - \overline{y}_{...}) + (\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...}) + (y_{ijk} - \overline{y}_{ij.})$$

son ortogonales entre sí. Podemos representar simbólicamente esta descomposición como

$$SCT = SCA + SCB + SCAB + SCR\ ,$$

donde:

- SCT denota la **Suma de Cuadrados Total**, que como en casos anteriores representa la variabilidad total observada de la respuesta alrededor de su media muestral global. En tanto la suma de los términos $y_{ijk} - \overline{y}_{...}$ ha de ser 0, esta suma tiene $N-1$ grados de libertad;
- SCA denota la **Suma de Cuadrados del Factor A** (o primer factor), que representa la parte de la variabilidad total que es explicada por el efecto aditivo de este factor. Nótese que cada término en la suma es $\overline{y}_{i..} - \overline{y}_{...} = \hat{\alpha}_i$, por lo que solo habrá $a-1$ términos independientes, y por tanto esta suma tiene $a-1$ grados de libertad;
- SCB denota la **Suma de Cuadrados del Factor B** (o segundo factor), que representa la variabilidad explicada por el efecto aditivo de este segundo factor. Similarmente a la SCA, cada término en la suma es $\overline{y}_{.j.} - \overline{y}_{...} = \hat{\beta}_j$, y por tanto esta suma tiene $b-1$ grados de libertad ;
- SCAB denota la **Suma de Cuadrados de la Interacción** entre los factores A y B, que representa la variabilidad explicada a través de los términos de interacción del modelo. Los términos de la suma son $\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...} = (\alpha\beta)_{ij}$, y por tanto se tendrán $(a-1)(b-1)$ grados de libertad asociados a esta suma;
- SCR denota la **Suma de Cuadrados Residual**, que representa la parte de la variabilidad total que no es explicada por ninguno de los términos anteriores, aglutinando por tanto

la variabilidad debida al error experimental, y contra la que se ha de comparar la significatividad de las partes de variabilidad explicada por los términos anteriores. Los términos en la suma son precisamente los residuos $y_{ijk} - \overline{y}_{ij\cdot} = e_{ijk}$, por lo que, como ya se ha mencionado antes, esta suma tendrá $ab(n-1)$ grados de libertad.

De modo similar, a través de la igualdad

$$y_{ijk} - \mu = (\overline{y}_{\cdots} - \mu) + (\overline{y}_{i\cdot\cdot} - \overline{y}_{\cdots}) + (\overline{y}_{\cdot j\cdot} - \overline{y}_{\cdots}) + (\overline{y}_{ij\cdot} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}_{\cdots}) + (y_{ijk} - \overline{y}_{ij\cdot})$$

es posible aplicar el teorema de Cochran para concluir que, bajo la hipótesis

$$H_0(\alpha, \beta, (\alpha\beta)) : \alpha_i = 0 \ \forall i;\ \beta_j = 0 \ \forall j;\ (\alpha\beta)_{ij} = 0 \ \forall i, j$$

de no influencia de ninguno de los factores y de no interacción, **las sumas de cuadrados anteriores se distribuyen como variables aleatorias chi-cuadrado con los grados de libertad indicados**, esto es:

$$\frac{\text{SCA}}{\sigma^2} \sim \chi^2_{a-1}, \quad \frac{\text{SCB}}{\sigma^2} \sim \chi^2_{b-1}, \quad \frac{\text{SCAB}}{\sigma^2} \sim \chi^2_{(a-1)(b-1)} \quad \text{y} \quad \frac{\text{SCR}}{\sigma^2} \sim \chi^2_{ab(n-1)}$$

Además, todas estas sumas de cuadrados serán independientes entre sí.

Esto permite entonces asegurar que los diferentes **cocientes entre sumas de cuadrados partidas por sus grados de libertad (esto es, los cocientes de las correspondientes medias de cuadrados) seguirán distribuciones *F* con los grados de libertad asociados,** lo que proporciona el mecanismo estadístico para contrastar las diferentes hipótesis de interés:

1. **Contraste sobre el factor A**: la hipótesis nula $H_0(\alpha) : \alpha_1 = \ldots = \alpha_a = 0$ afirma que el primer factor no tiene influencia aditiva sobre la respuesta, y se contrasta mediante el estadístico

$$F_0(\alpha) = \frac{SCA/(a-1)}{SCR/[ab(n-1)]} = \frac{MCA}{MCR} \overset{H_0(\alpha)}{\sim} F_{a-1,ab(n-1)}$$

2. **Contraste sobre el factor B**: la hipótesis nula $H_0(\beta) : \beta_1 = \ldots = \beta_b = 0$ afirma que el segundo factor no tiene influencia aditiva sobre la respuesta, y se contrasta mediante el estadístico

$$F_0(\beta) = \frac{SCB/(b-1)}{SCR/[ab(n-1)]} = \frac{MCB}{MCR} \overset{H_0(\alpha)}{\sim} F_{b-1,ab(n-1)}$$

3. **Contraste de la interacción**: en este caso la hipótesis nula $H_0(\alpha\beta) : \alpha\beta_{11} = \ldots = \alpha\beta_{ab} = 0$ establece que no existe ninguna interacción entre los factores A y B, y se contrasta mediante el estadístico

$$F_0(\alpha\beta) = \frac{SCAB/[(a-1)(b-1)]}{SCR/[ab(n-1)]} = \frac{MCAB}{MCR} \overset{H_0(\alpha)}{\sim} F_{(a-1)(b-1),ab(n-1)}$$

La información asociada al análisis de la varianza bifactorial puede resumirse en la correspondiente tabla ANOVA, que suele proporcionarse en la mayoría de paquetes estadísticos habituales con una forma similar a la expuesta en la Figura 1.8.

FUENTE DE VARIABILIDAD	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	CUADRADOS MEDIOS	F_0	p-valor
FACTOR A	SCA	$a-1$	MCF	MCF/MCR	$P(F_{a-1,ab(n-1)} > F_0)$
FACTOR B	SCB	$b-1$	MCB	MCB/MCR	$P(F_{b-1,ab(n-1)} > F_0)$
INTERACCIÓN	SCAB	$(a-1)(b-1)$	MCAB	MCAB/MCR	$P(F_{(a-1)(b-1),ab(n-1)} > F_0)$
ERROR	SCR	$ab(n-1)$	MCR		
TOTAL	SCT	$abn-1$			

Figura 1.8: Tabla ANOVA del modelo bifactorial.