Universidad Nacional Autónoma de México

ESCUELA NACIONAL DE ESTUDIOS PROFESIONALES

"ACATLAN"

MODELOS DESBALANCEADOS EN EL DISEÑO DE EXPERIMENTOS

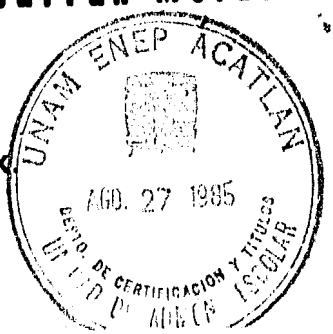
T E S I S

QUE PARA OBTENER EL TITULO DE:

A C T U A R I O

P R E S E N T A :

Reyna Albarrán Morales



ACATLAN, EDO. DE MEX. 1985

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

MODELOS DESBALANCEADOS EN EL DISEÑO DE EXPERIMENTOS

PRESENTACION

En el análisis de modelos lineales para el diseño de experimentos con datos balanceados existe un consenso general acerca de las hipótesis que se prueban bajo los encabezados de efectos -- principales e interacción. Sin embargo, los intentos que se -- han realizado para analizar los métodos empleados en el tratamiento de datos desbalanceados no han conducido a un acuerdo general. Esto aunado al desarrollo de paquetes estadísticos que facilitan el cálculo de las estadísticas de prueba sin que el - investigador tenga pleno conocimiento de las hipótesis que se - prueban, ha ocasionado gran confusión.

Por esta razón, el propósito de este trabajo es mostrar los problemas que presentan los modelos desbalanceados y exponer algunas de las soluciones que se han propuesto en la literatura estadística.

Para ello, en el capítulo I se introduce el concepto de experimento, de su diseño y de los modelos en el diseño de experimentos.

En el capítulo II se presenta un resumen del análisis de mode - los balanceados con uno y dos criterios de clasificación y del modelo en su forma general.

En el capítulo III se presenta un análisis de los modelos desbalanceados a partir de los resultados que se tienen para los balanceados con el fin de exhibir la problemática que existe.

En el capítulo IV se incluyen algunas contribuciones específicas en el análisis de modelos desbalanceados.

Es nuestra intención que estos cuatro capítulos sean de utilidad para aquellos investigadores que de alguna forma tengan que enfrentar el análisis de modelos desbalanceados.

I. INTRODUCCION

I.1 EL CONCEPTO DE EXPERIMENTO

Un experimento en términos generales, es un proceso planeado y controlado por medio del cual se incrementa la experiencia respecto algún fenómeno.

Al definir de este modo el experimento es evidente el alto rango de la actividad humana que tiene lugar a través de la experimentación.

Experimento es un vocablo derivado del latín experimentum, que es la acción y efecto de experimentar. Experimento es probar y examinar prácticamente la virtud y propiedades de una cosa. En ciencias físicas y naturales se entiende por experimento realizar operaciones destinadas a descubrir, comprobar y demostrar determinados fenómenos o principios científicos.

La palabra experimento suele ser empleada con distintas denotaciones, es decir, diferentes fijaciones conceptuales de su significado según el área o disciplina que se trate. En psicología se denomina experimento a la modificación planeada de condiciones con fines de observación; durante el experimento se modifica, por lo menos, una de las condiciones, la usualmente llamada variable independiente; la variable en dependencia de ésta, cuya variación es observada, se denomina variable dependiente o

de respuesta. En el curso del experimento, el resto de las condiciones necesarias se mantienen controladas en el caso de que se considere que pueden producir un efecto sobre la variable dependiente.

Experimento de Kaspar-Hauser. Este experimento se realiza criando a los animales recién nacidos aislados de sus padres y de otros animales. Se prueba la existencia de modos de conducta innatos. Según Freud, el instinto.

Desde el punto de vista de la pedagogía, todo nuevo método, plan y programa de estudio, así como los libros de texto, se consideran experimentales en el sentido de que los pedagogos que diseñan los métodos, los que elaboran los planes y programas de estudio y los que redactan libros, desconocen las reacciones de los estudiantes; por tanto esos materiales están en fase experimental hasta que se realiza la evaluación del plan, del programa o del texto con base en las reacciones de algunos alumnos.

En física, la experimentación ha sido y es de gran importancia ya que la simple observación del medio ambiente constituye la pauta del desarrollo de importantes experimentos que han dado origen a teorías de gran trascendencia. Ejemplo de ello se presenta en el experimento de Michelson (1881) el cual se montó -

por primera vez en 1881 para medir el movimiento de la tierra, - frente al hipotético éter. Como la tierra, en su movimiento de - translación alrededor del sol, se mueve con una gran velocidad - (30 km/s), la velocidad de propagación de la luz en un éter su--- puesto en reposo debería ser distinto según que la luz se propaga se en la dirección del movimiento de la tierra o en una dirección perpendicular a la misma. Del experimento de Michelson, llevado a cabo en el interferámetro de Michelson, se obtuvo el resultado de que la luz se propaga con la misma velocidad en todas direcciones. El experimento de Michelson dió pie para la teoría de la relatividad de Einstein.

Conforme con lo anterior, experimento es un proceso controlado - que se utiliza en diversas áreas o disciplinas para adquirir al-- gún conocimiento adicional de un tema de interés.

Cabe enfatizar que a través de la experimentación ha sido posible el desarrollo de nuevas teorías; se han comprendido algunos fenómenos y se han propuesto soluciones a problemas psíquicos, sociales, biológicos, químicos y físicos entre otros.

De acuerdo con los intereses del investigador, el experimento puede realizarse en ocasiones, para comparar dos o más conjuntos de datos que procedan de unidades que son tratados de manera diferente; o bien, simplemente para observar el comportamiento que su--- fren las unidades al aplicarles un determinado tratamiento.

Es importante señalar que en numerosas ocasiones las observaciones que se obtienen en un proceso de experimentación son producidas por fenómenos que en las condiciones que se observan se consideran aleatorios y por tal razón tienen que ser analizados mediante procedimientos que tomen en cuenta la incertidumbre asociada para establecer las conclusiones de una forma apropiada. En estas circunstancias, la estadística desempeña un papel fundamental tanto en el Diseño como en el Análisis de Experimentos.

I.2 EL DISEÑO DE UN EXPERIMENTO

Dada la importancia del experimento como parte del proceso de adquisición de conocimiento, es esencial su planeación de modo que su estructura y estrategia reúnan la más amplia información acerca del problema en estudio.

El diseño de un experimento es la secuencia completa de pasos programados de antemano, para asegurar que los datos producidos se obtendrán de tal forma que permitirán un análisis que conduzca a deducciones válidas con respecto al problema establecido con el mejor aprovechamiento de los recursos disponibles.

Antes de realizar el diseño de un experimento deben tenerse claros los objetivos de éste y considerar en general los recursos de que se dispone.

A través del diseño de un experimento se determina:

I) El número de observaciones que se van a realizar

II) El tipo de variables involucradas

III) La medición más adecuada de las observaciones

IV) El orden en que van a ser tomadas las observaciones

V) Los errores sistemáticos que se cometen

VI) El tipo de análisis que se aplicará a la información producida por el experimento

Al elaborar el diseño de un experimento es recomendable tomar en cuenta los siguientes aspectos:

Relatividad. Se refiere al balance que debe existir entre la capacidad económica, intelectual, de trabajo del experimentador, o del equipo de experimentadores, y el valor del problema de acuerdo a los objetivos perseguidos.

Globalización. Es la presentación sintética del problema antes de su análisis, llevado a cabo, con la colaboración de un grupo, usualmente interdisciplinario.

Sistematización. Es la ordenación adecuada de los problemas exigidos por los intereses y necesidades del experimentador.

Oportunidad. Se refiere al lugar, tiempo, situación económica, política, social, etc. propios para la ejecución del experimento.

Continuidad. Previendo todas las etapas del trabajo planeado, desde el inicial hasta el final.

Flexibilidad. Que permita posibles reajustes durante el experimento, sin quebrantar su unidad ni su continuidad.

Precisión y claridad en el diseño. De tal forma que un grupo interdisciplinario pueda llegar a conclusiones certeras del experimento.

A continuación se presentarán los pasos que se recomienda efectuar en el diseño de un experimento; pero antes, es necesario introducir algunos términos usuales:

Unidad Experimental. Es la mínima unidad a la cual se le aplica un sólo conjunto particular de condiciones experimentales.

Tratamiento. Es el conjunto particular de condiciones experimentales que se aplican a una unidad experimental dentro de los confines del diseño seleccionado.

Factor. Es una variable generalmente bajo control, cuyos efectos en los resultados experimentales son objeto del estudio.

Niveles de un factor. Son las diferentes modalidades que presenta un factor.

Efecto. Es el cambio que sufre la variable, respuesta que se observa y registra en el experimento, como consecuencia a cambios de nivel en un factor.

Pasos por seguir en el diseño de experimentos. De acuerdo con Kempthorne 1952, el diseño de un experimento que se analiza estadísticamente consta de los siguientes pasos:

1) Enunciado del problema
2) Formulación de objetivos
3) Proposición de la técnica experimental y el diseño
4) Examen de los sucesos posibles y referencias en que se basan las razones para la indagación que asegure que el experimento proporciona la información requerida en la extensión adecuada.

5) Consideración de los posibles resultados desde el punto de vista de los procedimientos estadísticos que se les aplicará, para asegurar que se satisfagan las suposiciones necesarias para que sean válidos estos procedimientos.

6) Ejecución del experimento

7) Aplicación de las técnicas estadísticas a los resultados experimentales.

8) Formulación de conclusiones. Deberá darse cuidadosa consideración a la validez de las conclusiones para la población de objetivos o eventos a la cual se van a aplicar.

9) Valuación de la investigación completa, particularmente con otras investigaciones del mismo problema o similares.

La comunicación entre el estadístico y el investigador es de vital importancia en el diseño de un experimento para este efecto, es de gran utilidad que el estadístico presente al investigador una lista de comprobación para planear un experimento. Una lista de este tipo fue preparada por Bicking y se reproduce, en esencia, a continuación:

A. Obtener un enunciado claro del problema

1. Identificar la nueva e importante área del problema.
2. Delinear el problema específico dentro de sus limitaciones usuales.
3. Definir el propósito exacto del programa de prueba.
4. Determinar la relación del problema particular con la investigación total o programa de desarrollo.

B. Reunir la información básica disponible

1. Investigar todas las fuentes de información disponibles.

C. Diseñar el programa de prueba

1. Sostener una conferencia con todas las partes concernientes:

   a) Enunciar los objetivos del experimento.

   b) Establecer el grado de confiabilidad que se considere conveniente.

   c) Esbozar las alternativas posibles de los sucesos.

   d) Determinar el rango práctico de estos factores y los niveles específicos a los que se harán pruebas.

   f) Escoger las mediciones finales que van a hacerse.

g) Considerar el efecto de la variabilidad en caso de realizar un muestreo y de la precisión del - método de prueba.

h) Considerar las posibles interacciones de los - factores.

i) Determinar las limitaciones de tiempo, costo, - materiales, recursos humanos, instrumentación y factores que sin ser de interés puedan modificar los resultados.

j) Considerar los aspectos de las relaciones humanas del programa.

2. Diseñar el programa en forma preliminar

a) Preparar un programa sistemático completo

b) Proporcionar las etapas de ejecución o adaptación del programa si es necesario.

c) Eliminar los efectos de las variables que no - están en estudio.

d) Elegir el método de análisis estadístico

e) Hacer las indicaciones prudentes para una acumulación ordenada de datos.

3. Revisar el diseño en colaboración con las partes concernientes

   a) Ajustar el programa de acuerdo con los comentarios.

   b) Desglosar en términos precisos los pasos a seguir.

D. Planear y llevar a cabo el trabajo experimental

1. Desarrollar métodos, materiales y equipo

2. Aplicar los métodos y técnicas

3. Supervisar y verificar los detalles; modificando el método si es necesario.

4. Registrar cualquier modificación al diseño del programa.

5. Recolectar el avance del programa.

E. Interpretar los resultados

1. Considerar todos los datos observados

2. Limitar las conclusiones o deducciones estrictas a partir de la evidencia obtenida.

3. Probar mediante experimentos independientes las controversias que susciten los datos.

4. Llegar a conclusiones tomando en consideración tanto la interpretación de los resultados así como su grado de confiabilidad.

5. Especificar lo que implican los resultados para su aplicación y para trabajos posteriores.

6. Tomar en cuenta todas las limitaciones impuestas por los métodos usados.

7. Enunciar los resultados y su confiabilidad.

G. Preparar el informe

1. Describir claramente el trabajo dando antecedentes, aclaraciones pertinentes del problema y del significado de los resultados.

2. Usar métodos gráficos y tabulares para la presentación de los datos en forma eficiente para usos futuros.

3. Suministrar información suficiente para que el lector pueda verificar resultados y sacar sus propias conclusiones.

4. Limitar las conclusiones a un resumen tal, que el trabajo evidencie su uso para consideraciones rápidas y acciones decisivas.

Para finalizar, es menester insistir en que el proceso de experimentación, para ser adecuado debe de configurarse como una actividad metódica y orientada por propósitos definidos. Los dos grandes males que debilitan y a la vez nulifican este proceso - son:

- La improvisación desordenada
- La rutina y la falta de objetivos

El mejor remedio contra esos grandes males es el diseño; éste - aumenta la confianza en el progreso metódico y bien calculado - del experimento hacia los objetivos definidos.

El diseño es, adicionalmente, un recurso para el control administrativo del proceso de experimentación.

El propósito del diseño es:

a) Que exista la mínima probabilidad de que se cometan errores.

b) Minimizar el costo, tiempo y esfuerzo invertido en el proceso de experimentación.

c) Evitar confusiones debido a las mediciones.

d) Considerar los posibles problemas que puedan surgir durante la experimentación.

e) Poder realizar un análisis que conduzca a una conclusión confiable sin hacer uso de métodos muy complicados.

I.3 MODELOS EN EL DISEÑO DE EXPERIMENTOS

Los modelos de diseño de experimentos son modelos matemáticos que utilizan métodos estadísticos para expresar los cambios existentes en la variable respuesta como consecuencia de la variación controlada de los factores en estudio y de los estímulos que están fuera del control del experimentador.

Su representación matemática más general es la siguiente:

$$Y = F\ (A, B, C, D, E, \ldots)$$

en donde Y representa la variable de respuesta, A,B,C,D,E,... representa la multitud de factores que inciden en dicha respuesta y $F(\cdot)$ es una función que describe esta relación. Usualmente, de los factores involucrados sólo un reducido número de ellos, por ejemplo A, B y C, son de interés y se encuentran bajo control. En estas condiciones el efecto total de los factores se descompone, bajo el supuesto de aditividad, en el debido a los factores bajo control y en el debido a los no controlados que en general se considera aleatorio.

Así, la forma original puede modificarse de la siguiente manera:

$$Y = f\ (A, B, C) + e\ (D, E, \ldots)$$

en donde f es una función que describe el efecto de A, B y C en la variable Y mientras que e (·) representa el efecto del resto de los factores que se considera como un error de naturaleza -- aleatoria.

Debido a la facilidad que representa para el análisis del modelo y tomando en cuenta la amplia gama de situaciones que pueden describirse de esa forma, f (A, B, C) es normalmente representado por una función lineal de las variables independientes. En el caso en que los atributos o factores bajo estudio sean de naturaleza cualitativa, es claro que esta representación lineal pueda obtenerse directamente mediante el empleo de variables indicadoras de los niveles de cada factor que se emplean para producir las observaciones. De esta forma el modelo original puede ser descrito de la siguiente manera:

$$Y = \sum_{s=1}^{q} \beta_s X_s + e$$

En el caso de modelos para el análisis de información experimental es también usual adoptar una notación complementaria que facilita la identificación de las piezas de información disponibles.

Por ejemplo, si en un experimento se utilizan dos factores bajo control y se realizan varias observaciones para una combinación de estos factores, es usual asignar a cada observación Y tres - subíndices de tal forma que Y_{ijk} representa la k-ésima observación realizada cuando los factores bajo control se fijan en los niveles i y j respectivamente.

De esta manera se obtiene la expresión:

$$Y_{ijk} = \sum_{s=1}^{q} \beta_s X_s + e_{ijk} \; ; \qquad \begin{matrix} i = 1,\ldots, a \\ j = 1,\ldots, b \\ k = 1,\ldots, nij \end{matrix} \quad \ldots\ (I.3.1)$$

En este modelo los parámetros involucrados son, en general, desconocidos y precisamente las hipótesis que usualmente son de interés, pueden ser planteadas en términos de afirmaciones sobre el valor que forman algunas de sus combinaciones lineales.

De esta forma, el objetivo original del experimento puede ser reducido a investigar la magnitud de las mencionadas combinaciones lineales que deben ser estimadas a partir de la información disponible.

La naturaleza de los errores implica, directamente la aleatoriedad en las observaciones en la variable de respuesta y las técnicas estadísticas para el análisis de este tipo de modelos ha-

cen uso de una serie de suposiciones sobre la distribución de probabilidades de estas variables para efectuar las estimaciones y las pruebas de hipótesis relevantes.

Usualmente, se supone que los errores involucrados en el modelo son independientes y tienen una distribución normal de media cero y varianza constante, pero desconocida σ^2.

De lo anterior, el modelo (I.3.1) puede describirse como

$$Y_{ijk} = \sum_{s=1}^{q} \beta_s X_s + e_{ijk} \qquad (I.3.2)$$

con $e_{ijk} \sim N(0, \sigma^2)$, independientes

lo que implica que

$$Y_{ijk} \sim N\left(\sum_{s=1}^{q} \beta_s X_s, \sigma^2\right), \text{independientes}$$

en donde los parámetros son desconocidos pero fijos.

Vale la pena mencionar que en algunas ocasiones, no se incluyen en un experimento todos los niveles de los factores bajo estudio, sino que sólo se seleccionan al azar algunos de ellos para su análisis. En esas condiciones con el fin de generalizar los resultados, los efectos de los factores se consideran aleatorios y el modelo recibe el nombre de modelo de efectos aleatorios. En lo que resta de este trabajo, se tratarán exclusivamente los modelos de efectos no aleatorios o fijos.

Adicionalmente, los modelos de diseño se han clasificado en balanceados, si cada tratamiento es aplicado al mismo número de observaciones, y desbalanceados si al menos existe un tratamiento que no contenga el mismo número de observaciones que los demás. Esta clasificación es muy importante porque en general, dependiendo de si el diseño es balanceado, la forma de analizarlo puede presentar diferentes modalidades.

El propósito principal de los modelos de diseño de experimentos es el de probar igualdad de efectos entre los niveles de los factores, esto es, probar que independientemente del nivel que se emplee de un factor, la respuesta permanece inalterada en cuyo caso se dice que el factor en cuestión no tiene efecto en la respuesta.

Básicamente, el método estadístico para probar este tipo de hipótesis es el conocido como cociente de verasimilitudes (Mood & Graybill) que consiste en comparar los valores máximos de la función de verosimilitud en el modelo original, también conocido como completo y el del modelo que incorpora la hipótesis bajo prueba, conocido como reducido.

II. DISEÑOS BALANCEADOS

II. DISEÑOS BALANCEADOS

II.1 UN CRITERIO DE CLASIFICACION

Antes de analizar en detalle los modelos con un criterio de clasificación, se expone un ejemplo del tipo de problemas que pueden ser resueltos con estos modelos.

Supóngase que se quiere investigar el efecto de una preparación de insulina, aplicada a tres dosis diferentes, en el porcentaje de reducción de azúcar en la sangre de conejos. Para ello se eligieron $3n$ conejos que tuvieran características homogéneas tales como peso, sexo, herencias, vigor y nutrición previa. Se formaron tres grupos cada uno integrado por n conejos seleccionados aleatoriamente. A cada grupo se le aplicó una dosis diferente de insulina para saber si realmente se producían cambios significativos en la reducción de azúcar en la sangre de los conejos.

Los datos se pueden presentar de la siguiente forma:

Dosis				
A_1	Y_{11}	Y_{12}	...	Y_{1n}
A_2	Y_{21}	Y_{22}	...	Y_{2n}
A_3	Y_{31}	Y_{32}	...	Y_{3n}

Donde Y_{ij} es el porcentaje de reducción de azúcar en el j-ésimo conejo al que se aplica la i-ésima dosis.

El modelo para representar este esquema puede ser expresado como sigue:

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

en donde μ representa el valor medio global de la variable de respuesta, y α_1, α_2, α_3 representan las desviaciones, respecto a μ debidas a los niveles del factor bajo estudio, esto es el efecto de cada uno de los tres niveles.

La hipótesis que usualmente se quiere probar en este tipo de modelo es la igualdad de los efectos debidos a los diferentes tratamientos esto es:

$$Ho: \alpha_1 = \alpha_2 = \ldots = \alpha_a \quad vs.$$

$$Ho: \alpha_i \neq \alpha_k \text{ para alguna } i \neq k \quad i,k = 1,2,\ldots a$$

En el ejemplo $a = 3$

En caso de aceptarse Ho se dice que la dosis no tiene efecto en la variable de respuesta, esto es, en la reducción de azúcar.

El análisis de los modelos con un criterio de clasificación particularmente para probar la hipótesis de igualdad de efectos, se lleva a cabo como sigue. El modelo correspondiente en el caso balanceado es el siguiente:

$$Y_{ij} = \mu + \alpha_i + e_{ij} \qquad \ldots \qquad (II.1.1)$$

donde

Y_{ij} es la j-ésima observación del i-ésimo nivel $\quad j = 1,\ldots, n$ $\quad i = 1,\ldots, a$

μ es la media general

α_i es el efecto de la i-ésima clase o nivel del factor bajo estudio

e_{ij} es el término del error aleatorio debido a las características específicas de Y_{ij}

Adicionalmente, se supone que $e_{ij} \sim N(0, \sigma^2)$ y que los errores son independientes.

Como consecuencia se tiene que $Y_{ij} \sim N(\mu + \alpha_i, \sigma^2)$:
$i = 1,2,\ldots, a$; $j = 1,2,\ldots, n$ y son independientes.

Para efectuar la prueba de hipótesis Ho: $\alpha_1 = \alpha_2 = \ldots = \alpha_a$ se procede de acuerdo a los siguientes pasos (R. Hicks).

I) Maximización de la función de verosimilitud en el modelo - completo.

Se establece la función de verosimilitud

$$L\ (f, \mu, \alpha_i, \sigma^2) = (2\Pi\sigma^2)^{-\frac{an}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{a} \sum_{j=1}^{n} (Y_{ij} - \mu - \alpha_i)^2\right)$$

en donde se tiene que $t = LnL(\cdot)$ una función más simple de optimizar tiene la expresión:

$$t = Ln[L(f, \mu, \alpha_i, \sigma^2)] = -\frac{an}{2} Ln\ 2\Pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{a} \sum_{j=1}^{n} (Y_{ij} - \mu - \alpha_i)^2$$

En virtud de que t es una función diferenciable, se obtienen las derivadas parciales respecto a μ, α_i, σ^2 y se igualan a cero

$$\frac{\partial t}{\partial \mu} = -\frac{1}{2\sigma^2} \sum_{i=1}^{a} \sum_{j=1}^{n} 2(Y_{ij} - \mu - \alpha_i)\ (-1) = 0$$

$$\frac{\partial t}{\partial \alpha_i} = -\frac{1}{2\sigma^2} \sum_{j=1}^{n} 2(Y_{ij} - \mu - \alpha_i)\ (-1) = 0 \qquad i = \overline{1,a}$$

$$\frac{\partial t}{\partial \sigma^2} = -\frac{an}{2} \frac{2\Pi}{2\Pi\sigma^2} - \sum_{i=1}^{a} \sum_{j=1}^{n} (Y_{ij} - \mu - \alpha_i)^2 \frac{1}{2\sigma^4} = 0$$

Reescribiendo lo anterior, se obtienen las expresiones que se conocen como ecuaciones normales:

$$\sum_{i=1}^{a}\sum_{j=1}^{n} Y_{ij} - an\hat{\mu} - n\sum_{i=1}^{a}\hat{\alpha}_i = 0$$

$$\sum_{j=1}^{n} Y_{ij} - n\hat{\mu} - n\,\hat{\alpha}_i = 0 \qquad \ldots \text{ II.1.2}$$

$$\hat{\sigma}^2 = \sum_{i=1}^{a}\sum_{j=1}^{n} \frac{(Y_{ij} - \hat{\mu} - \hat{\alpha}_i)^2}{an}$$

Para resolver estas ecuaciones respecto a $\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_a$, $\hat{\mu}$ se tiene un sistema de a + 1 incógnitas pero de rango a . Debe de notarse que en los modelos de regresión que se analizan en una forma similar, el sistema de ecuaciones es de rango completo. Para poder resolver el sistema (II.1.2) con una solución única se establece usualmente y por facilidad, la condición adicional

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

De esta manera (II.1.2) se convierte en:

$$\sum_{i=1}^{a}\sum_{j=1}^{n} Y_{ij} - an\,\hat{\mu} = 0 \qquad \Rightarrow \hat{\mu} = \sum_{i=1}^{a}\sum_{j=1}^{n} \frac{Y_{ij}}{an} \equiv \bar{Y}..$$

$$\sum_{j=1}^{n} Y_{ij} - n\,\hat{\mu} - n\,\hat{\alpha}_i = 0 \qquad \Rightarrow \hat{\alpha}_i = \sum_{j=1}^{n} \frac{Y_{ij}}{n} - \bar{Y}.. \equiv \bar{Y}_{i.} - \bar{Y}..\ \ i=\overline{1,a}$$

y por lo tanto $\hat{\sigma}^2$ que se obtiene en función de $\hat{\alpha}_1, \ldots \hat{\alpha}_i, \ldots \hat{\alpha}_a$ y $\hat{\mu}$ tiene la siguiente expresión:

$$\hat{\sigma}^2 = \frac{1}{an} \sum_{i=1}^{a} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_{i.})^2$$

Es posible verificar que el valor de $\hat{\sigma}^2$, que determina unívocamente el valor del máximo de la función de verosimilitud, es invariante ante cualquier solución del sistema de ecuaciones normales.

II) Maximización de la verosimilitud en el modelo reducido incorporando la hipótesis:

$$Ho: \quad \alpha_1 = \alpha_2 = \ldots = \alpha_a$$

Bajo la hipótesis Ho puede suprimirse el índice de los efectos, ya que todos son iguales y el modelo correspondiente es simplemente:

$$Y_{ij} = \mu + \alpha + e_{ij}$$

$$Y_{ij} = \mu^* + e_{ij} \; ; \quad \mu^* = \mu + \alpha \qquad i = 1,2,\ldots,a; \quad j = 1,2,\ldots n$$

Por lo tanto

$$Y_{ij} \sim N(\mu^*, \sigma^2)$$

y la función de verosimilitud es

$$L(f, \mu^*, \sigma^2) = (2\pi\sigma^2)^{-\frac{an}{2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{a} \sum_{j=1}^{n} (Y_{ij} - \mu^*)^2\right\}$$

Procediendo de la misma forma que en el modelo completo se tiene que

$$t = \text{Ln}(L(f, \mu^*, \sigma^2)) = -\frac{an}{2}\text{Ln}(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i\sum_j (Y_{ij} - \mu^*)^2$$

se deriva parcialmente con respecto a μ^*, σ^2 y se iguala a cero.

$$\frac{\partial t}{\partial \sigma^2} = -\frac{an}{2}\frac{2\pi}{2\pi\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\sum_i\sum_j (Y_{ij} - \hat{\mu}^*)^2 = 0$$

Lo equivalente a

$$\sum_i\sum_j Y_{ij} - an\hat{\mu}^* = 0 \quad \Longrightarrow \quad \hat{\mu}^* = \bar{Y}..$$

$$\hat{\sigma}_0^2 = \sum_i\sum_j \left(\frac{Y_{ij} - \bar{Y}..}{an}\right)^2$$

III) Se establece la forma de la región de rechazo de acuerdo al método de cociente de verosimilitud. Esto es, se rechaza Ho sí y sólo sí

$$\Lambda = \frac{\sup_{Ho} L(f, \mu^*, \sigma_0^2)}{\sup_{HoUH_1} L(f, \mu, \alpha_i, \sigma^2)} < k$$

con un valor $k > 0$ adecuadamente seleccionado, esto es, tal que satisfaga que $P(\text{rechazar Ho/Ho es cierto}) = \alpha$ para un valor de α cercano a cero (usualmente 0.05 ó 0.01).

Es conveniente observar que:

$$\frac{(2\pi\hat{\sigma}_0^2)^{-\frac{an}{2}}\ e^{-\frac{an}{2}}}{(2\pi\hat{\sigma}^2)^{-\frac{an}{2}}\ e^{-\frac{an}{2}}} < k$$

$$\Leftrightarrow \left[\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\right]^{-\frac{an}{2}} < k$$

$$\Leftrightarrow \left[\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right]^{\frac{an}{2}} < k$$

$$\Leftrightarrow \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} < k'$$

En este punto basta encontrar la distribución del cociente $\hat{\sigma}^2 / \hat{\sigma}_0^2$ bajo la hipótesis Ho para que el evento de rechazar la igualdad de efectos se lleve a cabo con la probabilidad preestablecida. Sin embargo, no es fácil determinar tal distribución y resulta conveniente reexpresar la forma de la región de rechazo en términos de una estadística cuya distribución sea conocida. Esto puede llevarse a cabo de la siguiente manera:

$$\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} < k'$$

$$\Leftrightarrow \frac{\sum_i \sum_j (Y_{ij} - \hat{\mu} - \hat{\alpha}_i)^2}{\sum_i \sum_j (Y_{ij} - \hat{\mu}^*)^2} < k' \quad \ldots \text{(II.1.3)}$$

En donde, el numerador de este cociente puede interpretarse como la suma de cuadrados del error debido al ajuste del modelo - completo (SCE_{MC}) mientras que el denominador, de forma análoga, resulta la suma de cuadrados del error debido al ajuste del modelo reducido (SCE_{MR}). Más aún se puede verificar que debido a que el modelo completo involucra un número mayor de paráme -- tros, y por tanto presenta mayor flexibilidad para el ajuste, se tiene que

$$SCE_{MC} < SCE_{MR}$$

De esta forma y como SCE_{MC} es mayor o igual a cero se puede definir.

$$SCE_{Ho} = SCE_{MR} - SCE_{MC} \geq 0$$

Es fácil comprobar que

$$SCE_{Ho} = \sum_{i=1}^{a} n(\bar{Y}_{i.} - \bar{Y}..)^2$$

De lo anterior y de (II.1.3) se tiene que:

$$\frac{SCE_{MC}}{SCE_{MR}} < k'$$

$$\Longleftrightarrow \quad \frac{SCE_{MC}}{SCE_{MC} + SCE_{Ho}} < k'$$

$$\Longleftrightarrow \quad \frac{1}{1 + \dfrac{SCE_{Ho}}{SCE_{MC}}} < k'$$

$$\Longleftrightarrow \quad \frac{SCE_{Ho}}{SCE_{MC}} > k''$$

$$\Longleftrightarrow \quad \frac{\sum_{i=1}^{a} n(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}{\sum_{i=1}^{a}\sum_{j=1}^{n} (Y_{ij} - \bar{Y}_{i\cdot})^2} > k''$$

En virtud de que puede demostrarse que SCE_{MC}/σ^2 se distribuye como una variable aleatoria χ^2 central con na-a grados de libertad mientras que:

$$\frac{SCE_{Ho}}{\sigma^2}$$

se distribuye como una variable aleatoria χ^2 no central con a-1 grados de libertad y parámetro de no centralidad $\lambda \geq 0$, en donde $\lambda = 0 \Longleftrightarrow$ Ho es cierta, y como además puede probarse que SCE_{MC} es independiente de SCE_{Ho} se tiene que:

$$F = \frac{SCE_{Ho}}{SCE_{MC}} \sim F_{(a-1,\, an-a)} \qquad \text{cuando Ho es cierta}$$

Así, F es la estadística adecuada para probar las hipótesis

$$Ho: \quad \alpha_1 = \alpha_2 \ldots = \alpha_a \qquad vs$$

$$Ha: \quad \alpha_i \neq \alpha_j \quad \text{para alguna } i \neq j$$

Naturalmente, se rechaza la hipótesis Ho si F es mayor que el cuantil apropiado de una distribución F con a-1 y na-a grados de libertad.

Usualmente, las magnitudes relevantes para probar la hipótesis de interés se presentan en la siguiente forma que se conoce como Tabla de Análisis de Varianza.

FUENTES DE VARIACION	GRADOS DE LIBERTAD	SUMA DE CUADRADOS	CUADRADOS MEDIOS	F
Tratamientos	$a-1$	$\sum_{i=1}^{a} n(\bar{Y}_{i.}-\bar{Y}_{..})^2$	$\sum_{i=1}^{a} \frac{n(\bar{Y}_{i.}-\bar{Y}_{..})^2}{a-1}$	$F=\frac{\sum_{i} \frac{n(\bar{Y}_{i.}-\bar{Y}_{..})^2}{a-1}}{\hat{\sigma}^2}$
Error	$a(n-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{i.})^2$	$\hat{\sigma}^2=\sum_{i=1}^{a}\sum_{j=1}^{n}\frac{(Y_{ij}-\bar{Y}_{i.})^2}{a(n-1)}$	
Total corregido por la media	$na-1$	$\sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{..})^2$		

Vale la pena notar que en la tabla anterior se tiene la pro - piedad, por construcción, de que las sumas de cuadrados de los dos primeros renglones suman al tercero.

II.2 DOS CRITERIOS DE CLASIFICACION SIN INTERACCION

Los modelos con dos criterios de clasificación se aplican en - aquellos casos donde se tienen dos factores de interés cuyos efectos en la variable de respuesta se pretende determinar. En general cada uno de los niveles de cada factor es combinado con todos los niveles del otro.

El concepto de interacción se emplea para describir la situación en la cual los efectos de uno de los factores dependen -- del particular nivel que adopte el otro factor. Los modelos - sin interacción se emplean cuando de antemano se sabe que no existe interacción entre los factores o no es de interés considerarla.

El análisis matemático de estos modelos en el caso balanceado resulta ser una generalización muy natural del que se efectúa en los modelos con un criterio de clasificación.

DEFINICION DEL MODELO

El modelo balanceado cuando no se considera la interacción, es el siguiente:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk} \qquad \begin{matrix} i = 1,\ldots,a \\ j = 1,\ldots,b \\ k = 1,\ldots,n \end{matrix} \quad \ldots \text{(II.2.1)}$$

En donde:

- μ es la media general
- α_i es el efecto del i-ésimo tratamiento del factor α
- β_j es el efecto del j-ésimo tratamiento del factor β
- e_{ijk} es el error aleatorio

Naturalmente se supone que:

$$e_{ijk} \sim N(0, \sigma^2) \qquad \begin{matrix} i = 1,2,\ldots,a \\ j = 1,2,\ldots,b \\ k = 1,2,\ldots,n \end{matrix}$$

y estos errores son independientes.

De igual manera que en la sección II.1 se obtienen las ecuaciones normales para el modelo (II.2.1)

$$abn\,\hat{\mu} + bn \sum_{i=1}^{a} \hat{\alpha}_i + an \sum_{j} \hat{\beta}_j = Y...$$

$$bn\,\hat{\mu} + bn\,\hat{\alpha}_i + n \sum_{j} \hat{\beta}_j = Y_{i..} \quad \forall\, i = \overline{1,a} \qquad ...(II.2.2)$$

$$an\,\hat{\mu} + n \sum \hat{\alpha}_i + an\,\hat{\beta}_j = Y_{.j.} \quad \forall\, j = \overline{1,b}$$

y $$\hat{\sigma}^2 = \frac{1}{abn} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (Y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j)^2$$

De nuevo (II.2.2) es un sistema de a + b + 1 ecuaciones con - - a + b + 1 incógnitas y rango a + b - 1 . Para poder resolver el - sistema (II.2.2) se deben establecer dos condiciones adicionales que usualmente se eligen como :

$$\sum_{i=1}^{a} \hat{\alpha}_i = 0 \qquad \sum_{j=1}^{b} \hat{\beta}_j = 0$$

y se obtiene la siguiente solución

$$\hat{\mu} = \bar{Y}...$$

$$\hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}...$$

$$\hat{\beta}_j = \bar{Y}_{.j.} - \bar{Y}...$$

$$\frac{SCE}{abn} = \hat{\sigma}^2 = \frac{1}{abn} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (Y_{ijk} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}...)^2$$

HIPOTESIS QUE SE PRUEBAN

Para probar Ho: $\beta_1 = \beta_2 = \ldots = \beta_b$ hay que considerar el modelo - reducido:

$$Y_{ijk} = \mu + \alpha_i + \beta + e_{ijk} = \mu^* + \alpha_i + e_{ijk}\ ;\ \mu^* = \mu + \beta$$

que es semejante al modelo (II.1.1). Consecuentemente,

$$SCE_{MR} = \sum_{i=1}^{a}\sum_{k=1}^{n} (Y_{ijk} - \bar{Y}_{i..})^2 \qquad y$$

$$SCE_{Ho} = SCE_{MR} - SCE_{MC} = \sum_{j=1}^{b} an\,(\bar{Y}_{.j.} - \bar{Y}...)^2$$

Como $e_{ijk} \sim N(0, \sigma^2)$ independientes, $\frac{SCE_{MC}}{\sigma^2}$ se distribuye como una χ^2 no central con b-1 grados de libertad y un parámetro de no centralidad que de nuevo es cero sí y sólo sí Ho es cierta. En virtud de la independencia de estas dos sumas de cuadrados se puede verificar que:

$$F_1 = \frac{SCE_{Ho}/(b-1)}{SCE_{MC}/(abn-a-b+1)} \sim F\,(b-1,\ abn-a-b+1) \text{ bajo Ho.}$$

F_1 entonces, es la estadística adecuada para probar la hipótesis Ho: $\beta_1 = \beta_2 = \ldots = \beta_b$.

De igual forma si se quiere probar la hipótesis

$$Ho: \alpha_1 = \alpha_2 = \ldots = \alpha_a$$

El modelo reducido por la hipótesis es:

$$Y_{ijk} = \mu + \alpha + \beta_j + e_{ijk} = \mu' + \beta_j + e_{ijk} \; ; \; \mu' = \mu + \alpha \quad (II.2.4)$$

Con

$$SCE_{MR} = \sum_j \sum_k (Y_{ijk} - \bar{Y}_{.j.})^2$$

y

$$SCE_{Ho} = SCE_{MR} - SCE_{MC} = b \sum_i \sum_k (\bar{Y}_{i..} - \bar{Y}_{...})^2$$

Y en forma análoga se obtiene que la estadística de prueba para esta hipótesis es la siguiente:

$$F_2 = \frac{b \sum_j \sum_k (\bar{Y}_{i..} - \bar{Y}_{...})^2 \; / \; (a - 1)}{\sum_i \sum_j \sum_k (Y_{ij} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2 / \; abn - a - b + 1}$$

Para este modelo, la información indispensable para efectuar la prueba de las dos hipótesis de interés puede presentarse conjuntamente en una sola tabla de análisis de varianza que es la siguiente:

FUENTE DE VARIACION	GRADOS DE LIBERTAD	SUMA DE CUADRADOS
Entre Tratamientos α_i	$a - 1$	$\sum_i bn(\bar{Y}_{i..} - \bar{Y}...)^2$
Entre Tratamientos β_j	$b - 1$	$\sum_j an(\bar{Y}_{.j.} - \bar{Y}...)^2$
Error	$abn - a - b + 1$	$\sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}...)^2$
Total Corregido por la media	$abn - 1$	$\sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}...)^2$

En esta tabla, es interesante notar que nuevamente resulta que la suma de cuadrados del renglón que corresponde al modelo doblemente reducido al incorporar las dos hipótesis se obtiene como la suma de las tres sumas de cuadrados superiores. Este hecho se puede interpretar en el sentido de que el error o la variabilidad de las observaciones respecto a su media global se puede descomponer en tres partes; una debido a los términos β_j, otra debida a los términos α_i una más al término e_{ij}.

Adicionalmente, el hecho de que las pruebas de los efectos de los factores puedan realizarse de manera independiente se conoce como la propiedad de ortogonalidad de efectos. Esta estructura que resulta muy conveniente no se presenta en todos los modelos de experimentos y es debida, como se ilustrará posteriormente, al balanceo de las observaciones.

II.3 UN MODELO GENERAL

Hasta el momento se han tratado los modelos con uno y dos criterios de clasificación balanceados. En esta sección se estudiará el modelo con efectos fijos que generaliza a los anteriores y permite el estudio de más de dos criterios de clasificación e inclusive modelos con interacción haciendo uso de la notación matricial. Se expondrán algunas de las características básicas del modelo y las hipótesis que se prueban usualmente.

Considérese el siguiente modelo:

$$\underline{Y} = X\underline{b} + \underline{e} \qquad \ldots (II.3.1)$$

donde

- $\underline{Y}$ es un vector de observaciones de la variable de respuesta de dimensión $n \times 1$
- $\underline{b}$ es un vector de parámetros de dismensión $p \times 1$
- X es una matriz de orden $n \times p$ formando ceros y unos que representan ausencia y presencia de los diferentes tratamientos con rango $r < p < n$
- $\underline{e}$ es un vector de errores aleatorios tal que

$$E(\underline{e}) = \underline{0}; \quad Var(e) = \sigma^2 I_n$$ [1/]

1/ Para vectores aleatorios se denotará por Var (•) a la correspondiente matriz de varianzas y covarianzas.

$y \quad \underline{e} \sim N(\underline{0}, \sigma^2 In)$. Por lo tanto, se tiene que $E(Y) = Xb$, $var(Y) = \sigma^2 In \quad Y \sim N(X\underline{b}, \sigma^2 In)$.

En este modelo, para efectuar la prueba de hipótesis sobre el vector de parámetros $\underline{b}$ usualmente se emplea el método de cociente de verosimilitudes que requiere de la maximización de la verosimilitud tanto en el modelo original o completo como en el modelo reducido al incorporar la hipótesis.

Para el caso del modelo completo se tiene que:

$$L(f, \underline{b}, \sigma^2) = (2\Pi\sigma^2)^{-\frac{n}{2}} \exp\{-\frac{1}{2\sigma^2}(Y - X\underline{b})'(Y - X\underline{b})\}$$

de donde el logaritmo de esta función está dado por

$$Ln[L(f, \underline{b}, \sigma^2)] = -\frac{n}{2} Ln(2\Pi\sigma^2) - \frac{1}{2\sigma^2}(Y - X\underline{b})'(Y - X\underline{b})$$

Derivando parcialmente con respecto a $\underline{b}$ y σ^2 e igualando a cero obtendremos

$$\frac{\partial}{\partial \underline{b}} Ln[L(f, \underline{b}, \sigma^2)] = -\frac{1}{2\sigma^2}(2X'X\hat{\underline{b}} - 2X'\underline{Y}) = 0$$

$$\frac{\partial}{\partial \sigma^2} Ln[L(f, \underline{b}, \sigma^2)] = -\frac{n}{\sigma^2} + \frac{1}{\sigma^4}(\underline{Y} - X\hat{\underline{b}})'(Y - X\hat{\underline{b}}) = 0$$

y por tanto

$$X'X\hat{\underline{b}} = X'\underline{Y} \qquad \ldots \text{(II.3.2)}$$

$$\hat{\sigma}^2 = \frac{1}{n}(Y'Y - \hat{\underline{b}}\,X'Y) \qquad \ldots \text{(II.3.3)}$$

Donde (II.3.2) son las llamadas ecuaciones normales. Como X'X es de rango $r<p$, X'X es singular y por tanto no existe una solución única ($\hat{b}$) al sistema (II.3.2). Una alternativa para encontrar una solución es a través de la inversa generalizada de X'X que denotaremos con G. De esta manera:

$$\hat{\underline{b}} = GX'\underline{Y}$$

Se puede advertir que a medida que el modelo (II.3.1) incluye un número mayor de factores y sus interacciones, es más difícil calcular la inversa generalizada. Por otro lado debe hacerse notar que aún cuando $\hat{\underline{b}}$ no es única, el estimador de $\underline{Y}$ ($\hat{\underline{Y}}$) así como la suma de cuadrados del error (SCE), son invariantes a la elección de la solución, ya que

$$\hat{Y} = X\hat{\underline{b}} = XGX'Y$$

Pero se puede verificar que XGX' es invariante a cualquier elección de G. Por lo tanto se concluye que $\hat{\underline{Y}}$ es invariante.

Un resultado similar se tiene para la suma de cuadrados del -- error (SCE) del modelo (II.3.1). Considérese el error

$$e = \underline{Y} - X\hat{\underline{b}} = \underline{Y} - \hat{Y}$$

entonces,

$$\begin{aligned} SCE = e'e &= (Y - \hat{Y})' (Y - \hat{Y}) \\ &= (Y - XGX'Y)'(Y - XGX'Y) \\ &= Y'(I - XGX')(I - XGX')Y \\ &= Y'(I - XGX')Y \end{aligned} \qquad \ldots \quad (II.3.4)$$

con (I-XGX') idempotente y además invariante a la solución particular que se obtenga de las ecuaciones normales.

De (II.3.4) y (II.3.3) se observa que

$$\hat{\sigma}^2 = SCE / n$$

Además, de (II.3.4) se verificará que

$$SCE/\sigma^2 \sim \chi^2_{(n-r)} \qquad \ldots \quad (II.3.5)$$

HIPOTESIS QUE SE PRUEBAN

Como se mencionó anteriormente y como se ilustró mediante el -- ejemplo de la sección (II.1), el propósito fundamental del análisis de un experimento del tipo que se trata en este trabajo es probar igualdad de efectos; para esto debe tenerse presente que sólo deben de probarse hipótesis que involucren combinaciones lineales de los parámetros del modelo cuyos estimadores -- sean invariantes ante cualquier solución de las ecuaciones normales. Es decir, se requiere que las hipótesis a probar se formulen en términos de lo que se conoce como funciones lineales, linealmente estimables. Una función lineal de los parámetros es linealmente estimable si puede ser estimada insesgadamente por una combinación lineal del vector de observaciones.

Esto es

$\underline{q}'b$ es estimable si $\exists\ \underline{t}'\ \ni\ t'E(Y) = q'b$

o equivalentemente $E(\underline{t}'\underline{Y}) = q'b$

con

q un vector de dimensión $p \times 1$ y

t un vector de dimensión $n \times 1$

Supongamos que se quiere probar la hipótesis lineal $K'\underline{b} = 0$ -- con K' una matriz no singular de dimensión $(s \times p)$ y $K'\underline{b}$ linealmente estimable. En consecuencia debe existir T tal que:

$$K'_{(s \times p)} = T'_{(s \times n)} \; X_{(n \times p)}$$

Donde $\hat{K'b} = K'\hat{\underline{b}}$ es el mejor estimador linealmente insesgado - de $K'\underline{b}$. Además

$$\begin{aligned} \text{var}\,(K'\hat{\underline{b}}) &= K' \text{ var}\,(\hat{b})\, K \\ &= K' \text{ var}\,(GX'Y)\, K \\ &= K' GX'XG' K\sigma^2 \\ &= K' GX'XG'X' T\sigma^2 \\ &= K' GK\sigma^2 = T' XGX' T\sigma^2 \end{aligned}$$

invariante.

En consecuencia

$$K'\hat{b} \sim N(K'b,\ K' GK\sigma^2)$$

lo que implica que

$$Q = \frac{(K'\hat{b})' (K' GK)^{-1} (K'\hat{b})}{\sigma^2} \sim \chi^2_{(s,\lambda)} \qquad \ldots \quad (III.3.6)$$

Obsérvese que el parámetro λ conocido como parámetro de no centralidad es igual a cero sí y sólo sí la hipótesis $K'\underline{b} = 0$ es cierta, ya que

$$\lambda = \frac{E(K'\hat{b})' \ (K'GK)^{-1} E(K'\hat{b})}{2\sigma^2} = \frac{(K'b)' \ (K'GK)^{-1} (K'b)}{2\sigma^2} = 0$$

$$\Leftrightarrow \quad K'b = 0$$

Se puede demostrar que Q y SCE son independientes y por lo tanto de (II.3.5) y (II.3.6) se obtiene que

$$F = \frac{Q/s}{SCE/N-r} \sim F_{(s,n-r)}$$

Esta estadística es la que se obtiene al aplicar el método de cociente de verosimilitudes generalizado para la prueba de $Ho: K'\underline{b} = 0$ vs. $H_1: K'\underline{b} \neq \underline{0}$. La forma de la región de rechazo es la siguiente:

$$C = \{\underline{Y} \ / \ F > K\}$$

en donde la constante K, igual que en los casos ya descritos satisface que P (rechazar Ho/Ho es cierta) = α con un valor preestablecido de α.

III. DISEÑOS DESBALANCEADOS

III. DISEÑOS DESBALANCEADOS

Al diseñar un experimento se procura facilitar su análisis y a su vez obtener resultados confiables. Por esta razón, entre -- otras cosas, siempre se procura que los datos estén balanceados es decir, que exista el mismo número de observaciones en cada - celda o para cada combinación de tratamientos. Sin embargo, en ocasiones, las restricciones en cuanto a la disponibilidad - del material experimental no permite asignar el mismo número de observaciones en cada tratamiento.

En el capítulo anterior fueron presentados con relativo detalle los modelos con uno y dos criterios de clasificación con datos balanceados. Se puede verificar que no se tienen problemas relacionados con la estimabilidad de las hipótesis involucradas, ni con la resolución de las ecuaciones normales.

En los capítulos siguientes se tratará acerca de las dificultades que suelen presentarse al trabajar con modelos desbalanceados.

III.1 UN CRITERIO DE CLASIFICACION

Los modelos desbalanceados, como se dijo anteriormente, se caracterizan porque el número de observaciones en las celdas (tratamientos) no es el mismo en todo el experimento. Denotaremos con n_i el número de observaciones de la i-ésima celda, de tal forma que el modelo con un criterio de clasificación desbalanceado se define como:

$$Y_{ij} = \mu + \alpha_i + e_{ij} \qquad \ldots \quad (III.1.1)$$

Con

Y_{ij} la j-ésima observación del i-ésimo nivel
$j = 1, \ldots, n_i$; $i = 1, \ldots, a$

μ es la media general

α_i el efecto del i-ésimo tratamiento

e_{ij} es el término del error aleatorio
$e_{ij} \sim N(0, \sigma^2)$ $\forall$ i, j independientes

La única diferencia entre (II.1.1) y (III.1.1) es el número de observaciones. Tomando esto en consideración veamos que sucede con las hipótesis que se prueban, si en lugar de realizar el análisis del modelo paso a paso como se hizo en la sección (II.1) se adopta la tabla de Análisis de Varianza de la misma sección cambiando:

n por n_i

an por $n_. = \sum_{i=1}^{a} n_i$

TABLA 3.1

FUENTE DE VARIACION	GRADOS DE LIBERTAD	SUMA DE CUADRADOS	CUADRADOS MEDIOS
Entre Tratamientos	$a - 1$	$\sum_{i=1}^{a} n_i \left(\frac{Y_{i.}}{n_i} - \frac{Y_{..}}{n_.} \right)^2$	$\sum_{i=1}^{a} n_i \left(\frac{Y_{i.}}{n_i} - \frac{Y_{..}}{n_.} \right)^2 / a-1$
Error	$n_. - a$	$\sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(Y_{ij} - \frac{Y_{i.}}{n_i} \right)^2$	$\sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(Y_{ij} - \frac{Y_{i.}}{n_i} \right)^2 / n_. - a$
Total corregido por la media	$n_. - 1$	$\sum_{i=1}^{a} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}..)^2$	

La estadística F tendrá la forma

$$F = \frac{\sum_{i=1}^{a} n_i \left(\frac{Y_{i.}}{n_i} - \frac{Y_{..}}{n_.} \right)^2 / a - 1}{\sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(Y_{ij} - \frac{Y_{i.}}{n_i} \right)^2 / n_. - a}$$

Es fácil verificar que esta es la estadística que se obtiene si se aplica el método de cociente de verosimilitudes en el modelo desbalanceado.

Otra forma de corroborar que la estadística descrita es la apropiada para probar la hipótesis de igualdad de efectos, consiste en establecer que:

I) $SCE_{Ho} / \sigma^2 \sim \chi^2_{(a-1,\,\lambda)}$; en donde el parámetro de no centralidad es cero sí y sólo sí la hipótesis Ho: $\alpha_1 = \alpha_2 = \cdots = \alpha_a$

II) $SCE_{MC} / \sigma^2 \sim \chi^2_{(n.-a)}$ independiente de SCE_{Ho} .

En particular respecto a I) se tiene que

$$SCE_{Ho} / \sigma^2 = \sum_{i=1}^{a} n_i \left(\frac{Y_{i.}}{n_i} - \frac{Y_{..}}{n_.} \right)^2 / \sigma^2$$

$$= \frac{Y' A Y}{\sigma^2} \qquad \ldots \quad (III.1.2)$$

Como

$$Y \sim N(Xb,\ \sigma^2 I); \text{ si } Z = \frac{Y}{\sigma} \quad Z \sim N\left(\frac{Xb}{\sigma},\ I\right)$$

Entonces

$$\frac{Y' A Y}{\sigma^2} = \frac{\sigma^2 Z' A Z}{\sigma^2} \sim \chi^2_{(r(A),\,\lambda)}$$

sí y sólo sí A es idempotente.

Dado que A se puede representar como:

$$
\begin{array}{c}
\\
n_1 \\
\vdots \\
n_1 \\
\\
n_2 \\
\vdots \\
n_2 \\
\vdots \\
n_a \\
\\
n_a
\end{array}
\left[\begin{array}{cccccccc}
\overbrace{}^{n_1} & & & & \overbrace{}^{n_2} & & \cdots & \overbrace{}^{n_a} \\
\frac{1}{n_1}-\frac{1}{n_\cdot} & \cdots & \frac{1}{n_1}-\frac{1}{n_\cdot} & & -\frac{1}{n_\cdot} & \cdots & -\frac{1}{n_\cdot} & \cdots & -\frac{1}{n_\cdot} & \cdots & -\frac{1}{n_\cdot} \\
\vdots & & & & & & & & & & \\
\frac{1}{n_1}-\frac{1}{n_\cdot} & \cdots & \frac{1}{n_1}-\frac{1}{n_\cdot} & & -\frac{1}{n_\cdot} & \cdots & -\frac{1}{n_\cdot} & \cdots & -\frac{1}{n_\cdot} & \cdots & -\frac{1}{n_\cdot} \\
-\frac{1}{n_\cdot} & \cdots & -\frac{1}{n_\cdot} & & \frac{1}{n_2}-\frac{1}{n_\cdot} & \cdots & \frac{1}{n_2}-\frac{1}{n_\cdot} & \cdots & -\frac{1}{n_\cdot} & \cdots & -\frac{1}{n_\cdot} \\
\vdots & & & & & & & & & & \\
-\frac{1}{n_\cdot} & \cdots & -\frac{1}{n_\cdot} & & \frac{1}{n_2}-\frac{1}{n_\cdot} & \cdots & \frac{1}{n_2}-\frac{1}{n_\cdot} & \cdots & -\frac{1}{n_\cdot} & \cdots & -\frac{1}{n_\cdot} \\
\vdots & & & & & & & & & & \\
-\frac{1}{n_\cdot} & \cdots & -\frac{1}{n_\cdot} & & -\frac{1}{n_\cdot} & \cdots & -\frac{1}{n_\cdot} & \cdots & \frac{1}{n_a}-\frac{1}{n_\cdot} & \cdots & \frac{1}{n_a}-\frac{1}{n_\cdot} \\
\\
-\frac{1}{n_\cdot} & \cdots & -\frac{1}{n_\cdot} & & -\frac{1}{n_\cdot} & \cdots & -\frac{1}{n_\cdot} & \cdots & \frac{1}{n_a}-\frac{1}{n_\cdot} & \cdots & \frac{1}{n_a}-\frac{1}{n_\cdot}
\end{array}\right]
$$

y que si se llama $B = A'A$, entonces

$$
b_{\substack{ij \\ (n_i \times n_j)}} \begin{cases} \frac{1}{n_i}-\frac{1}{n_\cdot} & \text{para } i=j \text{ con } i=1,\ldots,a \\ -\frac{1}{n_\cdot} & \text{para } i\neq j \text{ con } i,j=1,\ldots,a \end{cases}
$$

en consecuencia A es idempotente con rango a - 1. Lo que implica que

$$\frac{SCE_{Ho}}{\sigma^2} \sim \chi^2_{(a-1)}$$

con parámetro de no centralidad

$$\lambda = \frac{E(Y')AE(Y)}{2\sigma^2} \qquad \frac{b'X'AXb}{2\sigma^2} \longrightarrow$$

$$2\sigma^2\lambda = (\mu, \alpha_1, \alpha_2, \ldots \alpha_a) \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & n_1 - \frac{n_1^2}{n_\cdot} & -\frac{n_1 n_2}{n_\cdot} & \cdots & -\frac{n_1 n_a}{n_\cdot} \\ 0 & -\frac{n_2 n_1}{n_\cdot} & n_2 - \frac{n_2^2}{n_\cdot} & & -\frac{n_2 n_a}{n_\cdot} \\ \vdots & & & & \\ 0 & -\frac{n_a n_1}{n_\cdot} & -\frac{n_a n_2}{n_\cdot} & & n_a - \frac{n_a^2}{n_\cdot} \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \end{bmatrix}$$

$$= \sum_{i=1}^{a} \alpha_i^2 n_i - \frac{(\sum_{i=1}^{a} \alpha_i n_i)^2}{n_\cdot}$$

que es igual a cero sí y sólo sí Ho es cierta.

En esta sección se advirtió que trabajar con un criterio de clasificación desbalanceado no implica problema alguno. En el siguiente apartado se apreciará la dificultad que presentan los modelos desbalanceados más complejos.

III.2 DOS CRITERIOS DE CLASIFICACION SIN INTERACCION

Los modelos con dos criterios de clasificación sin interacción desbalanceados son semejantes al definido en la sección (II.2.1) con la única diferencia en el número de observaciones de cada - celda.

Esto es,

$$Y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk} \qquad \begin{matrix} i = 1,\ldots, a \\ j = 1,\ldots, b \\ k = 1,\ldots, n_{ij} \end{matrix} \quad \ldots \text{(III.2.1)}$$

donde n_{ij} es el número de observaciones a las que se les aplica el tratamiento (α_i , β_j).

Realizando un análisis semejante al de la sección anterior se - puede modificar la Tabla de Análisis de Varianza del apartado -- (II.2) para probar Ho: $\alpha_1 = \alpha_2 = \ldots = \alpha_a$ sustituyendo

n	por	n_{ij}
an	por	$\sum_i n_{ij} = n_{\cdot j}$
bn	por	$\sum_j n_{ij} = n_{i\cdot}$
abn	por	$\sum_i \sum_j n_{ij} = n_{\cdot\cdot}$

A diferencia del caso anterior, la estadística de prueba resultante no permite probar las mismas hipótesis. Esto es, no coincide con la que se obtiene cuando se aplica el método de cociente de verosimilitudes.

La Tabla de Análisis de Varianza que se produce con la sustitución es (3.2.1).

TABLA 3.2.1

FUENTE DE VARIACION	GRADOS DE LIBERTAD	SUMA DE CUADRADOS
Entre tratamientos	$a - 1$	$\sum_{i=1}^{a} n_{i.} (\bar{Y}_{i..} - \bar{Y}...)^2$
Entre tratamientos	$b - 1$	$\sum_{j=1}^{b} n_{.j} (\bar{Y}_{.j.} - \bar{Y}...)^2$
Error	$n.. - a - b + 1$	$\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}...)^2$
Total Corregido por la media	$abn - 1$	$\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}...)$

Con

$$Y_{i..} = \frac{Y_{i..}}{n_{i.}}$$

$$\bar{Y}_{.j.} = \frac{Y_{.j.}}{n_{.j}} \quad ; \quad \bar{Y}... = \frac{\bar{Y}...}{n..}$$

De esta manera, la estadística que se emplearía para probar la hipótesis

$$Ho: \quad \alpha_1 = \alpha_2 = \ldots = \alpha_a = \alpha$$

De acuerdo con (II.2) estaría dada por:

$$F^* = \frac{\sum_{i=1}^{a} n_{i.} (\bar{Y}_{i..} - \bar{Y}...)^2 / a-1}{\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (Y_{ijk} - Y_{i..} - Y_{.j.} + Y...)^2 / n.. - a - b+1}$$

Como ya se mencionó para que este sea apropiado para probar la hipótesis de interés, es necesario que el numerador dividido por σ^2 tenga una distribución χ^2 no central con un parámetro de no centralidad que sea igual a cero sí y sólo sí la hipótesis bajo prueba es cierta. En este caso se tiene que

$$\sum_{i=1}^{a} n_{i.} (\bar{Y}_{i..} - \bar{Y}...)^2 = Y'AY$$

Con la matriz A

$$
\begin{array}{c}
 \\
n_{1\cdot}\left\{\right. \\
 \\
n_{2\cdot}\left\{\right. \\
 \\
n_{a\cdot}\left\{\right.
\end{array}
\overbrace{\begin{bmatrix}
\frac{1}{n_{1\cdot}}-\frac{1}{n_{\cdot\cdot}} & \cdots & \frac{1}{n_{1\cdot}}-\frac{1}{n_{\cdot\cdot}} & -\frac{1}{n_{\cdot\cdot}} & \cdots & -\frac{1}{n_{\cdot\cdot}} \\
\vdots & & \vdots & \vdots & & \vdots \\
\frac{1}{n_{1\cdot}}-\frac{1}{n_{\cdot\cdot}} & \cdots & \frac{1}{n_{1\cdot}}-\frac{1}{n_{\cdot\cdot}} & -\frac{1}{n_{\cdot\cdot}} & \cdots & -\frac{1}{n_{\cdot\cdot}} \\
-\frac{1}{n_{\cdot\cdot}} & \cdots & -\frac{1}{n_{\cdot\cdot}} & \frac{1}{n_{2\cdot}}-\frac{1}{n_{\cdot\cdot}} & \cdots & -\frac{1}{n_{\cdot\cdot}} \\
\vdots & & \vdots & \vdots & & \vdots \\
-\frac{1}{n_{\cdot\cdot}} & \cdots & -\frac{1}{n_{\cdot\cdot}} & \frac{1}{n_{2\cdot}}-\frac{1}{n_{\cdot\cdot}} & \cdots & -\frac{1}{n_{\cdot\cdot}} \\
\vdots & & \vdots & \vdots & & \vdots \\
-\frac{1}{n_{\cdot\cdot}} & \cdots & -\frac{1}{n_{\cdot\cdot}} & -\frac{1}{n_{\cdot\cdot}} & \cdots & -\frac{1}{n_{\cdot\cdot}} \\
\vdots & & \vdots & \vdots & & \vdots \\
-\frac{1}{n_{\cdot\cdot}} & \cdots & -\frac{1}{n_{\cdot\cdot}} & -\frac{1}{n_{\cdot\cdot}} & \cdots & -\frac{1}{n_{\cdot\cdot}}
\end{bmatrix}}^{n_{1\cdot}}
\overbrace{\begin{bmatrix}
-\frac{1}{n_{\cdot\cdot}} & \cdots & -\frac{1}{n_{\cdot\cdot}} \\
\vdots & & \vdots \\
-\frac{1}{n_{\cdot\cdot}} & \cdots & -\frac{1}{n_{\cdot\cdot}} \\
\frac{1}{n_{2\cdot}}-\frac{1}{n_{\cdot\cdot}} & \cdots & \frac{1}{n_{2\cdot}}-\frac{1}{n_{\cdot\cdot}} \\
\vdots & & \vdots \\
\frac{1}{n_{2\cdot}}-\frac{1}{n_{\cdot\cdot}} & \cdots & \frac{1}{n_{2\cdot}}-\frac{1}{n_{\cdot\cdot}} \\
\vdots & & \vdots \\
-\frac{1}{n_{\cdot\cdot}} & \cdots & -\frac{1}{n_{\cdot\cdot}} \\
\vdots & & \vdots \\
-\frac{1}{n_{\cdot\cdot}} & \cdots & -\frac{1}{n_{\cdot\cdot}}
\end{bmatrix}}^{n_{2\cdot}}
\cdots
\overbrace{\begin{bmatrix}
-\frac{1}{n_{\cdot\cdot}} & \cdots & -\frac{1}{n_{\cdot\cdot}} \\
\vdots & & \vdots \\
-\frac{1}{n_{\cdot\cdot}} & \cdots & -\frac{1}{n_{\cdot\cdot}} \\
\vdots & & \vdots \\
-\frac{1}{n_{\cdot\cdot}} & \cdots & -\frac{1}{n_{\cdot\cdot}} \\
\vdots & & \vdots \\
\frac{1}{n_{a\cdot}}-\frac{1}{n_{\cdot\cdot}} & \cdots & \frac{1}{n_{a\cdot}}-\frac{1}{n_{\cdot\cdot}} \\
\vdots & & \vdots \\
\frac{1}{n_{a\cdot}}-\frac{1}{n_{\cdot\cdot}} & \cdots & \frac{1}{n_{a\cdot}}-\frac{1}{n_{\cdot\cdot}}
\end{bmatrix}}^{n_{a\cdot}}
$$

Se puede comprobar que A es idempotente.

Entonces como

$$Y \sim N (X\underline{b} , \sigma^2 I)$$

Haciendo un cambio de variable:

$$Z = \frac{Y}{\sigma} \quad \Rightarrow \quad Y = Z\sigma \ ;$$

$$Z \sim N\left(\frac{Xb}{\sigma}, I\right) \qquad \text{y en consecuencia}$$

$$\frac{Y'AY}{\sigma^2} = Z'AZ \sim \chi^2\left(r(A), \lambda\right)$$

Dado que el parámetro de no centralidad se puede calcular como sigue

$$\lambda = \left[E(Y')AE(Y)\right] / 2\sigma^2$$

$$\lambda = (b'X'AXb) / 2\sigma^2$$

Realizando algunas operaciones, se tiene que:

$$\lambda = \frac{b'}{2\sigma^2} R - Pb$$

donde

$$R_{(a+b+1)\times(a+b+1)} = \left|\begin{array}{c|ccc|ccc} n_{..} & n_{1.} & \cdots & n_{i.} \cdots n_{a.} & n_{.1} & \cdots & n_{.j} \cdots n_{.b} \\ n_{1.} & n_{1.} & & & & & \\ \vdots & & & & & & \\ n_{i.} & & n_{i.} & & & \{n_{ij}\} & \\ \vdots & & & & & & \\ n_{a.} & & & n_{a.} & & & \\ \hline n_{.j} & & & & & & \\ \vdots & & & & & & \\ n_{.j} & & \{n_{ji}\} & & & \{\sum_{k}^{a}\left(\frac{n_{ki}}{n_{k.}} n_{kj}\right)\} & \\ \vdots & & & & & & \\ n_{.b} & & & & & & \end{array}\right|$$

$$P_{(a+b+1)\times(a+b+1)} = \left|\begin{array}{c|c} n_{..}\ n_{1.}\ \cdots\ n_{i.}\ \cdots\ n_{a.} & n_{.1}\ \cdots\ n_{.j}\ \cdots\ n_{.b} \\ \vdots & \\ n_{1.} & \\ \vdots & \\ n_{i.} \quad \dfrac{n_{i.}n_{j.}}{n_{..}} & \dfrac{n_{i.}n_{.j}}{n_{..}} \\ \vdots & \\ n_{a.} & \\ \hline n_{.1} & \\ \vdots & \\ n_{.j} \quad \dfrac{n_{.i}n_{j.}}{n_{..}} & \dfrac{n_{.i}n_{.j}}{n_{..}} \\ \vdots & \\ n_{.b} & \end{array}\right|$$

o equivalentemente:

$$\lambda = \Big[\sum_i \alpha_i^2 \Big(n_{i.} - \frac{n_{i.}^2}{n_{..}}\Big)^2 - \sum_i \sum_{k\neq i} \alpha_i \alpha_k \frac{n_{i.}n_{k.}}{n_{..}} + 2 \sum_i \sum_j \alpha_i \beta_j \Big(n_{ij} - \frac{n_{i.}n_{.j}}{n_{..}}\Big) +$$

$$+ \sum_j \sum_i \beta_j^2 \Big(\frac{n_{ij}^2}{n_{i.}} - \frac{n_{.j}^2}{n_{..}} \Big) + \sum_j \sum_{h\neq j} \beta_j \beta_h \Big(\sum_i \frac{n_{ij}n_{ih}}{n_{i.}} - \frac{n_{.j}n_{.h}}{n_{..}} \Big) \Big] / 2\sigma^2 .$$

Bajo la hipótesis Ho : $\alpha_1 = \alpha_2 = \ldots = \alpha_a$

$$2\sigma^2\lambda = \alpha^2 \sum_i \Big(n_{i.} - \frac{n_{i.}}{n_{..}}\Big)^2 - \alpha^2 \sum_i \sum_{k\neq i} \frac{n_{i.}n_{k.}}{n_{..}} + 2\alpha \sum_i \sum_j \beta_j \Big(n_{ij} - \frac{n_{i.}n_{.j}}{n_{..}}\Big) +$$

$$\sum_j \beta_j^2 \Big(\sum_i \frac{n_{ij}^2}{n_{i.}} - \frac{n_{.j}^2}{n_{..}} \Big) + \sum_j \sum_{h\neq j} \beta_j \beta_h \Big(\sum_i \frac{n_{ij}n_{ih}}{n_{i.}} - \frac{n_{.j}n_{.h}}{n_{..}} \Big)$$

Por lo tanto $2\sigma^2\lambda$ no es en general idénticamente cero y en consecuencia la estadística F* no sirve para probar la hipótesis

$$Ho : \alpha_1 = \alpha_2 = \ldots = \alpha_a$$

Una manipulación de la expresión de λ permite mostrar que

$\lambda = 0$ sí y sólo sí

$$\alpha_i + \frac{1}{n_{i.}} \sum_{j=1}^{b} n_{ij}\beta_j - (\alpha_k - \frac{1}{n_{k.}} \sum_{j=1}^{b} n_{kj}\beta_j) = 0 \quad \forall \; i,k = 1,\ldots,a \quad k \neq i$$

de lo que se deriva que F* podría ser utilizada para probar la hipótesis

$$Ho: \quad \alpha_i + \frac{1}{n_{i.}} \sum_{j=1}^{b} n_{ij}\beta_j - (\alpha_k + \frac{1}{n_{k.}} \sum_{j=1}^{b} n_{kj}\beta_j) = 0$$

$$\forall \; i \neq k \quad i,k = 1,\ldots,a$$

que en general, no es una hipótesis de interés.

A través de lo expuesto en esta sección se puede concluír que el análisis de los modelos de diseño de experimentos desbalanceados no pueden ser tratados como una generalización de los modelos balanceados. Además en el caso balanceado F en (II.2) es la estadística adecuada para probar igualdad de efectos mientras que F* en el caso desbalanceado, sirve para probar igualdad de una combinación de efectos distinta.

Con el propósito de especificar la forma correcta de probar las hipótesis de interés, a continuación se desarrolla el análisis del modelo (III.2.1).

El estudio se efectúa utilizando notación matricial. Así, el modelo se describe como:

$$Y = Xb + e$$

con ecuaciones normales

$$X'X\hat{b} = X'Y$$

o bien

$$\begin{bmatrix} n_{..} & n_{1.} & \cdots & n_{a.} & n_{.1} & \cdots & n_{.b} \\ n_{1.} & n_{1.} & & & & & \\ \vdots & & & & (n_{ij}) & & \\ n_{a.} & & n_{a.} & & & & \\ n_{.1} & & & n_{.1} & & & \\ \vdots & & (n_{jj}) & & \ddots & & \\ n_{.b} & & & & & & n_{.b} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_a \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_b \end{bmatrix} = \begin{bmatrix} Y_{...} \\ Y_{1..} \\ \vdots \\ Y_{a..} \\ Y_{.1.} \\ \vdots \\ Y_{.b.} \end{bmatrix}$$

Como puede observarse la primera columna de X'X es igual a la suma de las siguientes a columnas así como a la suma de las últimas b por lo tanto el rango de X'X, está dado por $r(X'X) = 1 + a + b - 2 = a + b - 1$. Para resolver las ecuaciones normales Searle (1971) propone establecer dos restricciones adicionales. Particularmente, sugiere igualar dos elementos de $\hat{b}$ a cero.

Una de las formas más fáciles de proceder consiste en imponer - la restricción $\hat{\mu} = 0$ y $\hat{\alpha}_1 = 0$ ó bien $\hat{\beta}_b = 0$ de acuerdo si $a < b$ ó $a > b$. Cuando hay menos niveles de α que de β, es conve - niente para facilitar los cálculos, establecer $\hat{\alpha}_1 = 0$ y cuando son menos los niveles de β se establece $\hat{\beta}_b = 0$. Supongamos - que hay menos niveles en β que en α, por tanto establecemos - $\hat{\beta}_b = 0$ y $\hat{\mu} = 0$. En ese caso el sistema de ecuaciones se redu ce a la siguiente expresión:

$$\begin{bmatrix} n_{1\cdot} & & 0 & n_{11} & \cdots & n_{1,b-1} \\ & \ddots & & & & \\ 0 & & n_{a\cdot} & n_{a1} & \cdots & n_{a,b-1} \\ n_{11} & \cdots & n_{a1} & n_{\cdot 1} & & 0 \\ \vdots & & & & \ddots & \\ n_{1,b-1} & \cdots & n_{a,b-1} & 0 & & n_{\cdot,b-1} \end{bmatrix} \begin{bmatrix} \hat{\alpha}_1 \\ \\ \hat{\alpha}_a \\ \hat{\beta}_1 \\ \\ \hat{\beta}_{b-1} \end{bmatrix} = \begin{bmatrix} Y_{1\cdot\cdot} \\ \vdots \\ Y_{a\cdot\cdot} \\ Y_{\cdot 1\cdot} \\ \vdots \\ Y_{\cdot,b-1,\cdot} \end{bmatrix}$$

Resolviendo las primeras a ecuaciones

$$\hat{\alpha}_i = \bar{Y}_{i\cdot\cdot} - \frac{1}{n_{i\cdot}} \sum_{j=1}^{b-1} n_{ij}\, \hat{\beta}_j \qquad i = 1, \ldots, a$$

y sustituyendo estos valores en las últimas b-1 ecuaciones tene mos

$$n_{\cdot j} - \sum_{i=1}^{a} \frac{n_{ij}^2}{n_{i\cdot}}\, \hat{\beta}_j - \sum_{j' \neq j}^{b-1} \sum_{i=1}^{a} \frac{n_{ij}\, n_{ij'}}{n_{i\cdot}}\, \hat{\beta}_{j'} = Y_{\cdot j\cdot} - \sum_{i=1}^{a} n_{ij}\, \bar{Y}_{i\cdot\cdot}$$

para $j, j' = 1, 2, \ldots, b-1$

Similarmente, en forma vectorial tenemos $\underline{C}\hat{\underline{\beta}}_{b-1} = r$ con solución $\hat{\beta}_{b-1} = C^{-1} r$

donde $C = \{C_{jj'}\}$ y $r = \{r_j\}$ $\quad j = 1, \ldots, b-1$

$$\hat{\underline{\beta}}_{b-1} = \begin{bmatrix} \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_a \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{b-1} \end{bmatrix}$$

con $C_{jj} = n_{.j} - \sum_{i=1}^{a} \frac{n_{ij}^2}{n_{i.}}$ $\quad C_{jj'} = -\sum_{i=1}^{a} \frac{n_{ij}\, n_{ij'}}{n_{i.}}$ con $j \neq j'$

y $\quad r_j = Y_{.j.} - \sum_{i=1}^{a} n_{ij} \bar{Y}_{i..}$

Si denotamos

$$D_a = \begin{bmatrix} n_{1.} & & 0 \\ & n_{2.} & \\ 0 & & \\ & & n_{a.} \end{bmatrix}$$

Y

$$N_{ax(b-1)} = \begin{bmatrix} n_{11} & \cdots & n_{1\,b-1} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ n_{a1} & & n_{a\,,b-1} \end{bmatrix}$$

$$M_{ax(b-1)} = D_a^{-1} N = (n_{ij} / n_{i.}) \quad \text{con} \quad i=1,\ldots,a \quad j=1,\ldots,b-1$$

y $\quad \bar{Y}_a = D_a^{-1} Y_a = (\bar{Y}_{i.}) \quad \text{con} \quad i=1,\ldots,a$

y por tanto $\quad \hat{\underline{\alpha}} = D_a^{-1} \underline{Y}_a - M \hat{\underline{\beta}}_{b-1} = \bar{\underline{Y}}_a - M \hat{\underline{\beta}}_{b-1}$

De esta manera

$$\hat{\underline{b}} = \begin{bmatrix} 0 \\ \hat{\underline{\alpha}} \\ \hat{\underline{\beta}}_{b-1} \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \bar{\underline{Y}}_a - MC_r^{-1} \\ C_r^{-1} \\ 0 \end{bmatrix}$$

De aquí además es posible obtener la inversa generalizada de $X'X$ que da lugar a la solución $\hat{\underline{b}}$.

$$G = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & D_a^{-1} - MC^{-1}M' & -MC^{-1} & 0 \\ 0 & -C^{-1}M' & C^{-1} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Al obtener de esta forma los estimadores de $\underline{\alpha}$ y $\underline{\beta}$ así como la inversa generalizada, aparentemente no existe ninguna dificultad debido al desbalanceo, sin embargo, se advierte que en la práctica es complicado obtener la inversa de la matriz C.

ANALISIS DE VARIANZA

Para facilitar el análisis de varianza de modelos con más de un criterio de clasificación es usual el empleo del concepto de la reducción en sumas de cuadrados $R(\cdot)$.

Así, de acuerdo a la notación propuesta por Searle dentro del paréntesis se determinan los parámetros del modelo que se ajusta. De esta forma $R(\mu, \alpha, \beta) = \hat{b}\, X'Y$ es la reducción en la suma de cuadrados de $\underline{Y}$ debido a ajustar el modelo $Y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$. Por otra parte, la reducción en sumas de cuadrados por haber ajustado α dado que ya se había ajustado de antemano μ y β se denota por $R(\alpha/\mu, \beta)$ y se calcula como

$$R(\alpha/\mu, \beta) = R(\alpha, \mu, \beta) - R(\mu, \beta)$$

Ahora bien, para el modelo (III.2.1) si definimos

$$\underline{Y}'_{\beta} = (Y_{\cdot 1}, \ldots Y_{\cdot b-1}, \cdot) \quad \text{entonces}$$

$$R(\mu, \alpha, \beta) = (\bar{\underline{Y}}_a - MC^{-1} r)' Y_a + (C^{-1} r)' \underline{Y}_{\beta}$$

Como

$$r = \underline{Y}_{\beta} - M' \underline{Y}_a$$

$$R(\mu, \alpha, \beta) = \bar{Y}_a' Y_a + r' C^{-1} r$$

y además de (II.1)

$$R(\mu) = n.. \bar{Y}^2 ...$$

$$R(\mu, \alpha) = \sum_{i=1}^{a} n_{i.} \bar{Y}_{i..}^2 = \bar{Y}_a' Y_a$$

Por lo tanto

$$R(\mu, \alpha, \beta) = R(\mu, \alpha) + r' C^{-1} r$$

$$= R(\mu, \alpha) + \hat{\beta}' r$$

De la tabla (III.2.1) tenemos que la reducción en suma de cuadrados debido a α en presencia de μ está dada por:

$$\sum_{i=1}^{a} \frac{Y_{i.}^2}{n_{i.}} - \frac{Y_{...}^2}{n_{..}} = R(\mu, \alpha) - R(\mu) = R(\alpha/\mu)$$

Se observa que $R(\alpha/\mu)$ no involucra la presencia del parámetro β y por ello es posible calcular una reducción alternativa debido a α ahora en presencia también de β .

$$R(\alpha/\mu, \beta) = R(\mu, \alpha, \beta) - R(\mu, \beta)$$

$$= \bar{Y}_{a}' \, Y_{a} + r' C^{-1} r - \sum_{j=1}^{b} n_{.j} \bar{Y}_{.j}^2$$

Es fácil comprobar que:

$$R(\alpha/\mu) \neq R(\alpha/\mu, \beta)$$

Por lo tanto, existen dos formas en las que es factible descomponer $R(\alpha, \mu, \beta)$

i) Ajustando primero μ enseguida μ y α y por último μ, α y β.

ii) Ajustando primero μ enseguida μ y β y por último μ, α y β.

En correspondencia con estas descomposiciones se pueden obtener dos diferentes Tablas de Análisis de Varianza que se presentan a continuación.

TABLA 3.2.2

FUENTE DE VARIACION	GRADOS DE LIBERTAD	SUMA DE CUADRADOS
Media	1	$R(\mu) = n_{..}\bar{Y}^2_{...}$
α Dado μ	a - 1	$R(\alpha/\mu) = \sum_{i=1}^{a} n_{i.}\bar{Y}^2_{i..} - n_{..}\bar{Y}^2_{...}$
β Dado α y μ	b - 1	$R(\beta/\mu,\alpha) = r' C^{-1} r$
Error	N-a-b + 1	$SCE = \sum_{i}\sum_{j}\sum_{k=1}^{n_{ij}} Y^2_{ijk} - \sum_{i} n_{i.}\bar{Y}_{i..} - r'C^{-1}r$

TABLA 3.2.3

Media	1	$R(\mu) = n_{..}\bar{Y}^2_{...}$
β Dado μ	b - 1	$R(\beta/\mu) = \sum_{j=1}^{b} n_{.j}\bar{Y}^2_{.j.} - n_{..}\bar{Y}^2_{...}$
α Dado β y μ	a - 1	$R(\alpha/\mu,\beta) = \sum_{i=1}^{a} n_{i.}\bar{Y}^2_{i..} + r'C^{-1}r - \sum_{j=1}^{b} n_{.j}\bar{Y}^2_{.j.}$
Error	N-a-b+1	$SCE = \sum_{i}\sum_{j}\sum_{k} Y^2_{ijk} - \sum_{i} n_{i.}\bar{Y}_{i..} - r'C^{-1}r$

Es importante remarcar que en el caso balanceado se verifica de (II.2) y (II.1) que

$$R(\alpha/\mu,\beta) = R(\mu,\alpha,\beta) - R(\mu,\beta) = \frac{Y_{i\cdot\cdot}^2}{bn} - \frac{Y_{\cdot\cdot\cdot}^2}{abn}$$

$$R(\alpha/\mu) = R(\mu,\alpha) - R(\alpha) = \frac{Y_{i\cdot\cdot}^2}{bn} \quad \frac{Y_{\cdot\cdot\cdot}^2}{abn}$$

De modo que no existe ninguna diferencia debido al orden en el que se ajustan los parámetros. Este hecho sugiere que mientras que en el caso balanceado se pueden efectuar las pruebas de hipótesis sobre cada uno de los factores ignorando para el cálculo de las reducciones el factor restante, este no es el caso cuando se presenta una situación de desbalanceo.

De hecho en el caso balanceado se tiene que

$R(\mu,\alpha,\beta) = R(\mu) + R(\alpha/\mu) + R(\beta/\mu)$ expresión que es simétrica en α y β mientras que cuando se presenta el desbalanceo se tiene que

$$R(\mu,\alpha,\beta) = R(\mu) + R(\alpha/\mu) + R(\beta/\mu,\alpha)$$

o bien

$$R(\mu,\alpha,\beta) = R(\mu) + R(\beta/\mu) + R(\alpha/\mu,\beta) .$$

PRUEBAS DE HIPOTESIS

Como se estableció anteriormente $R(\alpha/\mu)$ sirve para probar la hipótesis

$$Ho: \quad \alpha_i + \frac{1}{n_{i.}} \sum_{j=1}^{b} n_{ij} \beta_j = \alpha_{i'} + \frac{1}{n_{i'.}} \sum_{j=1}^{b} n_{i'j} \beta_j \qquad \forall\ i \neq i'$$

y similarmente $R(\beta/\mu)$ / b-1 es el numerador de la estadística F que prueba la hipótesis

$$Ho: \quad \beta_j + \frac{1}{n_{.j}} \sum_{i=1}^{a} n_{ij} \alpha_i = \beta_{j'} + \frac{1}{n_{.j'}} \sum_{i=1}^{a} n_{ij'} \alpha_i \qquad \forall\ j \neq j'$$

En virtud de que las hipótesis que son de interés en el análisis de modelos de experimentos son usualmente:

$$Ho: \quad \alpha_i - \alpha_a = 0 \qquad \forall\ i = 1,2,\ldots, a-1$$

$$Ho: \quad \beta_j - \beta_b = 0 \qquad \forall\ j = 1,2,\ldots, b-1$$

se pretende comprobar si éstas pueden ser probadas a través de las reducciones $R(\alpha/\mu, \beta)$ y $R(\beta/\mu, \alpha)$ respectivamente.

Supóngase que

$$Ho: \beta_j - \beta_b = 0 \qquad \forall\, j = 1,2,\ldots b-1$$

es la hipótesis que se desea probar que puede ser expresada como

$$Ho: k'\underline{b} = 0 \quad \text{con} \quad k' = [\,0\, 1_{b-1}, 0, I_{b-1}, -1_{b-1}]$$

De la sección (II.3) se sabe que

$$Q/s = (k'\underline{\hat{b}})' (k' Gk)^{-1} k'\underline{\hat{b}}/s$$

es el numerador de la estadística F que prueba Ho: $k'\underline{b} = 0$ de tal forma que, para este caso, se tiene que mostrar que

$$Q = R(\beta/\mu, \alpha)$$

Para ello calculemos:

$$k' G = [\,0,\ -C^{-1}M',\ C^{-1}, 0\,]$$

$$k'Gk = C^{-1} \qquad y$$

$$k'\hat{b} = k' GX\ Y = (-C^{-1}M'Y_\alpha + C^{-1}Y_\beta)$$

Entonces

$$Q = (-C^{-1}M' Y_\alpha + C^{-1} Y_\beta)' (C^{-1})^{-1}(-C^{-1}M' Y_\alpha + C^{-1} Y_\beta)$$

$$= (Y_\beta - M' D_\alpha^{-1} Y_\alpha)' C^{-1} (Y_\beta - M' D_\alpha^{-1} Y_\alpha)$$

$$= r' C^{-1} r$$

$$= \hat{\beta}_r' = R(\beta/\mu, \alpha)$$

De igual forma se demuestra que $R(\alpha/\mu, \beta)$ sirve para probar la hipótesis

$$Ho: \alpha_i - \alpha_a = 0 \qquad \forall\ i = 1, \ldots, a-1$$

Para concluir debe subrayarse que al tener distinto número de observaciones en las celdas se pierde la ortogonalidad de efectos lo que ocasiona que

$$R(\alpha/\mu, \beta) \neq R(\alpha/\mu)$$

y

$$R(\beta/\mu, \alpha) \neq R(\beta/\mu)$$

Aún cuando se ha trabajado con modelos con dos criterios de clasificación sin interacción se advierte que existe dificultad en el análisis y que además se obtienen diferentes sumas de cuadrados que sirven para probar distintas hipótesis.

III.3 DOS CRITERIOS DE CLASIFICACION CON INTERACCION

Con el propósito de especificar lo que sucede cuando se trabaja con un modelo desbalanceado, en esta sección se resuelve de -- acuerdo con la metodología expuesta por Kendall [1966], un modelo desbalanceado con dos criterios de clasificación que incluye interacción.

Para formular el modelo lineal, de acuerdo con lo establecido - por Kendall, se requiere de un parámetro para la media de las - observaciones μ_{ij} en cada una de las ab celdas que puede ex - presarse en términos de:

$\mu_{..}$ una media común a todas las observaciones

$\mu_{i.}$ una media común a las observaciones en la i-ésima fila (al i-ésimo nivel de α)

$\mu_{.j}$ una media común a las observaciones en la j-ésima columna (el j-ésimo nivel de β)

De donde se observa que el modelo incluye $1 + a + b + ab$ paráme - tros, de los cuales sólo ab son linealmente independientes.

Kendall define a

$$\mu = \mu_{..}$$

$$\alpha_i = \mu_{i.} - \mu_{..}$$

$$\beta_j = \mu_{.j} - \mu_{..}$$

$$\gamma_{ij} = \mu_{ij} - (\mu_{i.} + \mu_{.j}) + \mu_{..}$$

y representa el modelo lineal como

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk} = \mu_{ij} + e_{ij} \quad \ldots \quad (III.3.1)$$

donde

μ es la media general

α_i es el efecto de la i-ésima fila $i = \overline{1, a}$

β_j es el efecto de la j-ésima columna $j = \overline{1, b}$

γ_{ij} es el efecto debido a la interacción entre la i-ésima fila y la j-ésima columna.

Imponiendo al modelo (III.3.1) las condiciones no estimables

$$0 = \sum_{i=1}^{a} n_{i.}\,\alpha_i = \sum_{j=1}^{b} n_{.j}\,\beta_j = \sum_{i=1}^{a} n_{ij}\,\gamma_{ij}$$

$$= \sum_{j} n_{ij}\,\gamma_{ij} = \sum_{i}\sum_{j} n_{ij}\gamma_{ij} \qquad \text{con} \quad \begin{matrix} i = \overline{1, a-1} \\ j = \overline{1, b-1} \end{matrix}$$

Se obtienen a + b + 1 parámetros que pueden ser expresados en términos de los otros de la siguiente forma:

$$\alpha_a = -\sum_{i=1}^{a-1} n_{i.}\,\alpha_i \mid n_{a.}$$

$$\beta_b = -\sum_{j=1}^{b-1} n_{.j}\,\beta_j \mid n_{.b}$$

$$\gamma_{aj} = -\sum_{i=1}^{a-1} n_{ij}\,\gamma_{ij} \mid n_{aj} \qquad j = \overline{1, b-1}$$

$$\gamma_{ib} = -\sum_{j=1}^{b-1} n_{ij}\,\gamma_{ij} \mid n_{ib} \qquad i = \overline{1, a-1}$$

$$\gamma_{ij} = \sum_{i=1}^{a-1}\sum_{j=1}^{b-1} n_{ij}\,\gamma_{ij} \mid n_{ab}$$

De esta forma la matriz de diseño X correspondiente al modelo lineal definido en (III.3.1) que involucra los parámetros eliminados en términos de los otros puede escribirse como se muestra en la figura (III.3.1). Premultiplicando X por su transpuesta, se encuentra que

$X'X =$

	1	(a-1)	(b-1)	(a-1)(b-1)
1	n	0	0	0
(a-1)	0	A	D	0
(b-1)	0	D'	B	0
(a-1)(b-1)	0	0	0	C

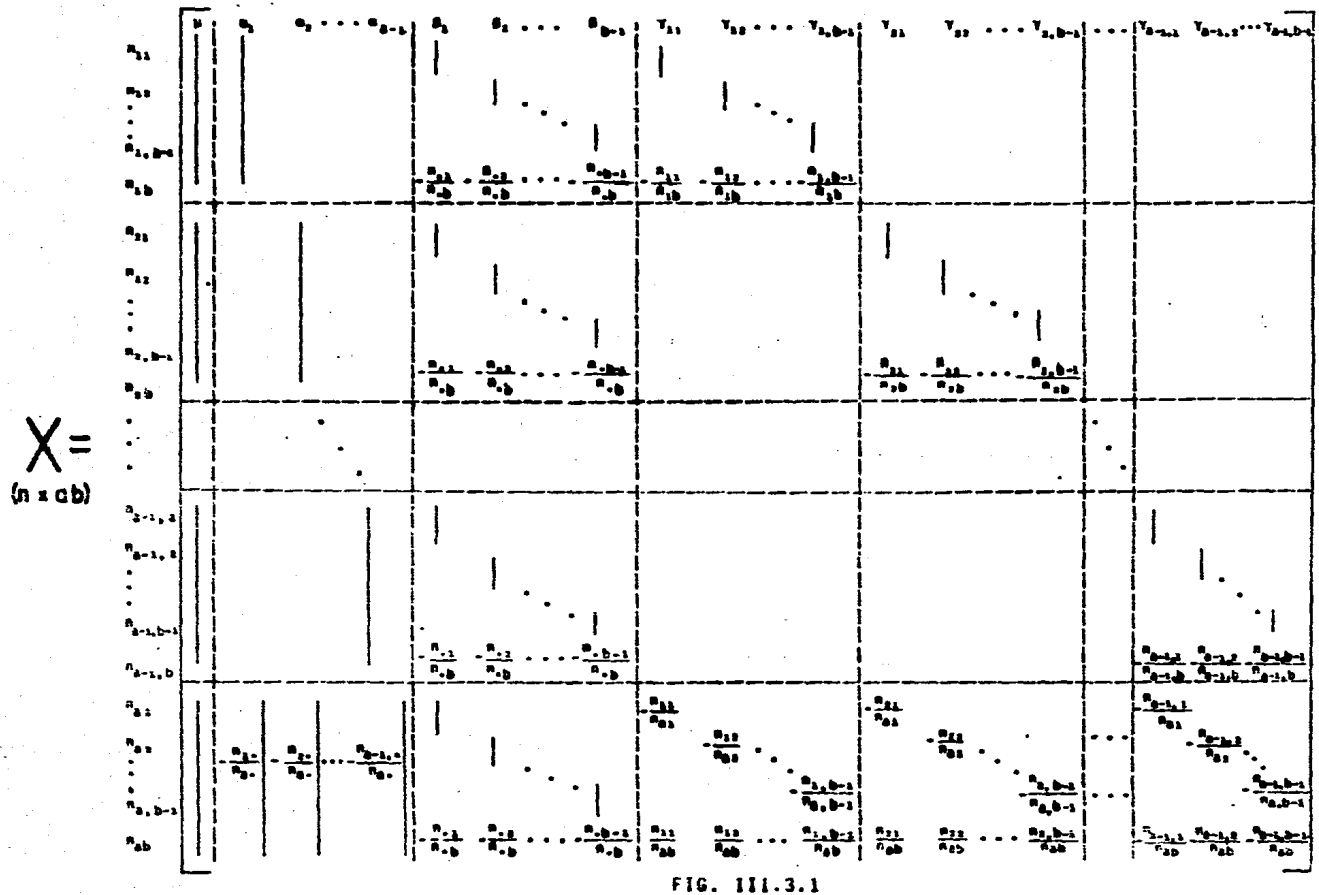


FIG. III.3.1

Donde:

$$A_{(a-1)x(a-1)} = \begin{bmatrix} n_{1\cdot} + \frac{n_{1\cdot}^2}{n_{a\cdot}} & \frac{n_{1\cdot}\, n_{2\cdot}}{n_{a\cdot}} & \cdots & \frac{n_{1\cdot}\, n_{a-1,\cdot}}{n_{a\cdot}} \\ & n_{2\cdot} + \frac{n_{2\cdot}^2}{n_{a\cdot}} & \cdots & \frac{n_{2\cdot}\, n_{a-1,\cdot}}{n_{a\cdot}} \\ & & \ddots & \\ & & & n_{a-1,\cdot} + \frac{n_{a-1,\cdot}^2}{n_{a\cdot}} \end{bmatrix}$$

$$B_{(b-1)x(b-1)} = \begin{bmatrix} n_{\cdot 1} + \frac{n_{\cdot 1}^2}{n_{\cdot b}} & \frac{n_{\cdot 1}\, n_{\cdot 2}}{n_{\cdot b}} & \cdots & \frac{n_{\cdot 1}\, n_{\cdot b-1}}{n_{\cdot b}} \\ & n_{\cdot 2} + \frac{n_{\cdot 2}^2}{n_{\cdot b}} & \cdots & \frac{n_{\cdot 2}\, n_{\cdot b-1}}{n_{\cdot b}} \\ & & \ddots & \\ & & & n_{\cdot b-1} + \frac{n_{\cdot b-1}^2}{n_{\cdot b}} \end{bmatrix}$$

$$C_{(kl),(mq)} = n_{kl} + n_{kl}^2 \left(\frac{1}{n_{kb}} + \frac{1}{n_{al}} + \frac{1}{n_{ab}}\right) \quad \text{SI } k=m,\ l=q$$

$$= n_{kl}\, n_{kq} \left(\frac{1}{n_{kb}} + \frac{1}{n_{ab}}\right) \quad \text{SI } k\neq m,\ l=q$$

$$= n_{kl}\, n_{mq} / n_{ab} \quad \text{SI } k\neq m,\ l\neq q$$

Donde (kl) representa el $\{(k-1)(b-1)+l\}$-ésimo renglón y (mq) la $\{(m-1)(b-1)+q\}$ - ésima columna de C.

En general X'X puede invertirse numéricamente, pero esta operación, aún cuando el modelo es sencillo, resulta difícil de realizar. Sin embargo, se verifica que si D = 0 la inversa de X'X se obtiene en forma sencilla

$$\begin{bmatrix} n & & & 0 \\ & A & & \\ & & B & \\ 0 & & & C \end{bmatrix}$$

con

$$(X'X)^{-1} = \begin{bmatrix} n^{-1} & & & 0 \\ & A^{-1} & & \\ & & B^{-1} & \\ 0 & & & C^{-1} \end{bmatrix}$$

Para que la matriz D sea igual a cero se requiere que

$$\frac{n_{ij}}{n_{.j}} = \frac{n_{ib}}{n_{.b}} \qquad i = \overline{1, a-1} \qquad j = \overline{1, b-1}$$

y que

$$\frac{n_{aj}}{n_{.j}} = \frac{n_{ab}}{n_{.b}}$$

El caso más simple se presenta cuando $n_{ij} = n_{i'j'} \quad \forall\ i,j,i',j'$ es decir cuando el modelo es balanceado. No obstante, es difícil que en la realidad se tenga esta situación por lo que en la mayoría de las ocasiones la matriz D es diferente de cero y en consecuencia

$$X'X = \begin{bmatrix} n & & & \\ & A & D & \\ & D' & B & \\ & & & C \end{bmatrix}$$

cuya inversa puede escribirse como

$$(X'X)^{-1} = \begin{bmatrix} n^{-1} & & & \\ & E & & \\ & & & C^{-1} \end{bmatrix}$$

con

$$E = \begin{bmatrix} A & D \\ D' & B \end{bmatrix}^{-1} = \begin{bmatrix} (A - DB^{-1}D')^{-1} & -(A - DB^{-1}D')^{-1}DB^{-1} \\ -(B - D'A^{-1}D)^{-1}D'A^{-1} & (B - D'A^{-1}D)^{-1} \end{bmatrix}$$

Al trabajar con datos desbalanceados los estimadores de los parámetros difieren de los que se obtienen en el caso balanceado. Lo que provoca que las sumas de cuadrados del modelo ajustado - así como las del error varíen.

Este hecho es muy importante ya que las hipótesis asociadas a - éstas también son diferentes a las que se prueban en el caso -- balanceado. Esta diferencia se verifica al descomponer la suma de cuadrados del modelo completo en

$$SC_{MC} = Y'Y - \hat{\beta}' X'Y = Y'Y - \hat{\beta}' X'X \hat{\beta}$$

$$= Y'Y - n\hat{\mu} + \begin{pmatrix} \hat{\underline{\alpha}}_i \\ \hat{\underline{\beta}}_j \end{pmatrix}' \begin{bmatrix} A & D \\ D' & B \end{bmatrix} \begin{pmatrix} \hat{\underline{\alpha}}_i \\ \hat{\beta}_j \end{pmatrix} + \hat{\underline{Y}}_{ij}\, C\, \hat{\underline{Y}}_{ij}$$

. . . (III.3.2)

donde

$\hat{\underline{\alpha}}_i$ es el vector que contiene los estimadores debido al efecto de fila

$\hat{\underline{\beta}}_j$ es el vector que contiene los estimadores debido al efecto de columna

$\hat{\underline{Y}}_{ij}$ es el vector que contiene los estimadores debido al efecto de interacción

De (III.3.2) puede observarse que los estimadores de los tres - grupos de parámetros: la media general, el efecto de filas y columnas y la interacción no están correlacionados entre los tres grupos pero sí están correlacionados dentro de ellos. Cuando - D = 0 no existe correlación entre el efecto de filas y el efecto de columnas lo que ocasiona que estos dos efectos sean ortogonales entre sí.

Para calcular, en el caso desbalanceado, las sumas de cuadrados debido al efecto de columnas y filas se considera el término medio de la ecuación (III.3.2). Sin embargo, para calcular por - ejemplo el efecto debido únicamente a las filas, existen dos -- formas diferentes:

1) Calcular la reducción en las sumas de cuadrados debido al ajuste del modelo completo y restarle las sumas de cuadrados debido al modelo que sólo excluye a $\hat{\underline{\alpha}}_i$.

2) Calcular la reducción en las sumas de cuadrados debido al ajuste del modelo completo y restarle la suma de cuadrados debida al modelo que exluyen a $\hat{\underline{\alpha}}_i$ y $\hat{\underline{\beta}}_j$

Se puede verificar que al igual que en el modelo con dos criterios de clasificación sin interacción desbalanceado, estas dos formas para calcular la contribución de $\hat{\underline{\alpha}}_i$ al modelo conducen a resultados diferentes, mientras que en el caso balanceado los resultados son iguales.

Searle [1971] presenta el Análisis de Varianza del modelo (II.3.1) basándose en las reducciones $R(\cdot)$, y propone dos diferentes Tablas de Análisis de Varianza de acuerdo al orden en el cual se ajustan los parámetros.

Asimismo, para determinar que hipótesis se prueban al comparar las reducciones de las sumas de cuadrados de las diferentes entradas en las Tablas de Análisis de Varianza con las sumas de cuadrados debido al error, Searle comprueba que:

$R(\alpha / \mu)$ es el numerador que sirve para probar la hipótesis

$$Ho: \alpha_i + \frac{1}{n_{i.}} \sum_j n_{ij} (\beta_j + \gamma_{ij}) = \alpha_{i'} + \frac{1}{n_{i'.}} \sum_j n_{i'j} (\beta_j + \gamma_{i'j}) \quad \forall\, i \neq i'$$

Similarmente $R(\alpha / \mu, \beta)$ sirve para probar la hipótesis

$$H : \alpha_i \left[n_{i.} - \sum_j n_{ij}^2 / n_{.j} \right] - \left[\sum_{i' \neq i} \sum_j n_{ij} n_{i'j} / n_{.j} \right] \alpha_{i'} +$$

$$\sum_j \gamma_{ij} \left[n_{ij} - n_{ij}^2 / n_{.j} \right] - \sum_{i' \neq i} \left[n_{ij} n_{i'j} / n_{.j} \right] \gamma_{i'j} = 0$$

De esta forma para probar la contribución del factor α es posible utilizar tanto $R(\alpha / \mu)$ o $R(\alpha / \mu, \beta)$.

Alrededor de este hecho se han desarrollado una serie de críticas sobre cúal es la suma de cuadrados adecuada para probar igualdad de efectos de filas o de columnas. Además, en virtud de que el análisis desbalanceado involucra cálculos complicados que aún disponiendo de los avances actuales en computación resultan poco manejables, algunos autores estudiosos de los mode-

los de diseño de experimentos, han sugerido métodos alternativos aproximados que facilitan el análisis, lo cual ha dado lugar a confusiones debido a que en ocasiones los métodos alternativos en su afán de simplificar el problema algebraico, olvidan tomar en consideración las hipótesis que realmente se están probando. En el próximo capítulo se presentan algunos trabajos que tratan este problema.

IV. ALGUNAS TECNICAS PARA EL ANALISIS DE DISEÑOS DESBALANCEADOS

IV. ALGUNAS TECNICAS PARA EL ANALISIS DE DISEÑOS DESBALANCEADOS

Kramer (1955) en su trabajo discute tres métodos que son usuales para obtener las sumas de cuadrados de los efectos principales y de interacción en los modelos desbalanceados con dos criterios - de clasificación.

Supóngase el modelo

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk} \qquad \begin{matrix} i = \overline{1,a} \\ j = \overline{1,b} \\ k = \overline{1,n_{ij}} \end{matrix} \qquad \text{(IV.1.1)}$$

Un análisis preliminar, que en general se realizá para estudiár el caso desbalanceado, consiste en considerar un modelo con un - solo criterio de clasificación, cuyos niveles quedan definidos - por todas las combinaciones de los niveles de los factores originales. De esta forma, para realizar pruebas estadísticas sobre igualdad de efectos, únicamente se separa la suma de cuadrados - entre clases (debido al factor) de la suma de cuadrados dentro - de clases (debido error). La tabla de Análisis de Varianza que se deriva es la siguiente:

FUENTE	GRADOS DE LIBERTAD	SUMA DE CUADRADOS
ENTRE SUBCLASES	$ab - 1$	$SCF = \sum_i \sum_j \frac{Y_{ij.}^2}{n_{ij}} - \frac{Y_{..}^2}{n_{..}}$
DENTRO SUBCLASES	$n_{..} - ab$	SCE Por diferencia
TOTAL	$n_{..} - 1$	$SCT = \sum_i \sum_j \sum_k Y_{ijk}^2 - \frac{Y_{...}^2}{n_{..}}$

En donde como es usual el cociente

$$F = \frac{SCF \,/\, ab - 1}{SCE \,/\, n_{..} - ab}$$

tiene una distribución F con ab-1 y $n_{..} - ab$ grados de libertad bajo la hipótesis nula de igualdad de tratamientos, y resulta la estadística adecuada para llevar a cabo la prueba correspondiente - aún cuando exista el desbalanceo.

Cuando en el análisis se pretende estudiar por separado los efectos principales y de interacción, las sumas de cuadrados asociadas a cada uno de los factores α y β, así como a la interacción,

se pueden, en teoría, calcular a través de la técnica general para modelos lineales descrita en la sección (II.3). Como se puede observar, esta técnica involucra cálculos que en general resultan muy laboriosos. Por esta razón, se han propuesto varios métodos para el cálculo de estadísticas de prueba, que en algunos casos producen estadísticas similares a las del cociente de verosimilitud, pero que en general son diferentes.

Kramer en su artículo discute dos métodos ya conocidos en la literatura : el de Ajuste de Constantes y el de Medias Ponderadas, y propone adicionalmente, el de Medias Ponderadas Modificado.

El método de Ajuste de Constantes, considera el autor, es el óptimo cuando se presume ausencia de interacción. Este método fue propuesto por Yates (1934) y consiste en ajustar un modelo a los datos, de tal forma que los téminos constantes determinen un conjunto de medias de clase, con la propiedad que la suma de cuadrados ponderados por las frecuencias de las clases, de la diferencia de estas medias y las medias observadas sea mínima.

Esto es, se minimiza

$$S = \sum_i \sum_j n_{ij} \left(\bar{Y}_{ij\cdot} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j \right)^2$$

y se obtienen las ecuaciones normales que coinciden con las expuestas en la sección (III.2).

Otro método que el autor comenta es el de Medias Ponderadas sugerido por Yates (1934) el cual a su juicio es el óptimo cuando existe interacción.

El método consiste en obtener medias por clase $X_{ij} = \sum_k \frac{Y_{ijk}}{n_{ij}}$ y considerar un modelo en términos de las medias marginales:

$$\bar{X}_{i.} = \sum_j \frac{X_{ij}}{b} \qquad \bar{X}_{.j} = \sum_i \frac{X_{ij}}{a}$$

De esta manera, del modelo original se tiene que:

$$\bar{X}_{i.} = \mu + \alpha_i + \bar{\beta}_. + \bar{\gamma}_{i.} + \bar{e}_{i..}$$

$$= \mu^* + \alpha_i^* + \bar{e}_{i..}$$

en donde:

$$\mu^* = \mu + \bar{\beta}_.$$

$$\alpha_i^* = \alpha_i + \bar{\gamma}_{i.}$$

$$\bar{\beta}_. = \sum_j \frac{\beta_j}{b}$$

$$\bar{\gamma}_{i.} = \sum_j \frac{\gamma_{ij}}{b}$$

$$\bar{e}_{i..} = \sum_j \sum_k \frac{e_{ijk}}{bn_{ij}}$$

Análogamente se obtiene :

$$\bar{X}_{\cdot j} = \mu^* + \beta^*_j + \bar{e}_{\cdot j}.$$

con

$$\mu^* = \mu + \bar{\alpha}.$$

$$\beta^*_j = \beta_j + \bar{\gamma}_{\cdot j}$$

$$\bar{\alpha}. = \sum_i \frac{\alpha_i}{a}$$

$$\bar{e}_{\cdot j}. = \sum_i \sum_k \frac{e_{ijk}}{an_{ij}}$$

Ahora bien, para probar igualdad de efectos del factor α se considera la hipótesis nula

$$Ho : \alpha^*_1 = \alpha^*_2 = \ldots = \alpha^*_a = \alpha^{**}$$

y de acuerdo al Capítulo II es necesario encontrar los estimadores de máxima verosimilitud del modelo reducido:

$$\bar{X}_{i\cdot} = \mu^{**} + \bar{e}_{i\cdot\cdot}$$

con

$$\mu^{**} = \mu^* + \alpha^{**}$$

Dado que

$$E(\bar{X}_{i\cdot}) = (\mu^{**}) \quad y \quad Var(\bar{X}_{i\cdot}) = \frac{\sigma^2_\alpha}{b^2} \sum_j \frac{1}{n_{ij}}$$

la función de verosimilitud está dada por:

$$f(\bar{X}_{1\cdot}, \bar{X}_{2\cdot}, \ldots, \bar{X}_{a\cdot}) = \prod_i \left(\frac{2\pi\sigma^2_\alpha}{b^2} \sum_j 1/n_{ij} \right)^{-\frac{1}{2}} \exp\left\{ \frac{-b^2}{2\sigma^2_\alpha \sum_j 1/n_{ij}} (\bar{X}_{i\cdot} - \mu^{**})^2 \right\}$$

Si se define

$$\omega_i = \left(\frac{1}{b^2} \sum_j 1/n_{ij} \right)^{-1}$$

entonces,

$$f(\bar{X}_{1.}, \bar{X}_{2.}, \ldots, \bar{X}_{a.}) = \prod_i^a \left[(2\Pi\sigma_\alpha^2 \omega_i)^{-\frac{1}{2}} \exp\left(-\frac{\omega_i}{2\sigma^2}(X_{i.} - \mu^{**})^2\right) \right]$$

tomando logaritmos se obtiene:

$$t = Lnf = \sum_i \left[-\frac{1}{2} Ln\left(2\Pi\sigma_\alpha^2 \omega_i^{-1}\right) - \frac{\omega_i}{2\sigma_\alpha^2}(\bar{X}_{i.} - \mu^{**})^2 \right]$$

derivando con respecto a μ^{**} y a σ_α^2 e igualando a cero

$$\frac{\partial t}{\partial \mu^{**}} = -\frac{2}{2\sigma_\alpha^2} \sum_i \omega_i (\bar{X}_{i.} - \hat{\mu}^{**})$$

de donde

$$\frac{\partial t}{\partial \mu^{**}} = 0 \Rightarrow \hat{\mu}^{**} = \sum_i \omega_i \bar{X}_{i.} / \sum_i \omega_i$$

$$\frac{\partial t}{\partial \sigma_\alpha^2} = -\frac{1}{2}\left[\frac{a}{\sigma_\alpha^2} - \frac{1}{\sigma_\alpha^4} \sum_i \omega_i (\bar{X}_{i.} - \hat{\mu}^{**})^2 \right]$$

de manera que,

$$\frac{\partial t}{\partial \sigma_\alpha^2} = 0 \Rightarrow \hat{\sigma}_\alpha^2 = \sum_i \frac{\omega_i (\bar{X}_{i.} - \hat{\mu}^{**})^2}{a}$$

sustituyendo el valor de $\hat{\mu}^{**}$ se obtiene la expresión

$$\sigma_\alpha^2 = \sum_i \omega_i (\bar{X}_{i.} - \sum_i \omega_i X_{i.} / \sum_i \omega_i)^2 / a$$

Puede demostrarse que bajo Ho, $a\hat{\sigma}_{\alpha}^{2}/\sigma$ tiene una distribución - $\chi^{2}_{(a-1)}$ independientemente de SCE y en consecuencia

$$F_{\alpha} = \frac{a\sigma_{\alpha}^{2}/a-1}{SCE/n..-ab} \quad \text{sigue una distribución } F_{(a-1),(n..-ab)}$$

Por lo tanto F_{α} es la estadística que se puede utilizar para probar igualdad de efectos del factor α^{*}.

De igual forma, para probar la hipótesis $Ho' = \beta_1^{*}=\beta_2^{*}=\ldots=\beta_b^{*}$ se tiene que la suma de cuadrados del modelo simplificado y reducido por la hipótesis es:

$$\hat{\sigma}_{\beta}^{2} = \sum_{j} \upsilon_j \bar{X}_{.j} - \left(\sum_{j=1}^{b} \upsilon_j \bar{X}_{.j}\right)^{2} / \sum_{j} \upsilon_j / b$$

con

$$\upsilon_j = \left(\frac{1}{a^{2}} \sum_{i} 1 / n_{ij}\right)^{-1}$$

y por lo tanto, una estadística de prueba para Ho' es la siguiente:

$$F_{\beta} = \frac{b\,\hat{\sigma}_{\beta}^{2} / b-1}{SCE/n..-ab}$$

Kramer considera que estas pruebas tienen poca potencia debido a - que $\bar{X}_{i.}$ y $\bar{X}_{.j}$ no son estimadores de varianza mínima.

El autor propone una variante del método de medias ponderadas en el cual las medias marginales se calculan de la siguiente forma:

$$\bar{X}'_{i.} = \frac{1}{n..} \sum_j n_{.j} X_{ij}$$

$$\bar{X}'_{.j} = \frac{1}{n..} \sum_i n_{i.} X_{ij}$$

En este método como en el anterior, la variación entre las medias $\bar{X}'_{i.}$ es independiente del efecto de β. Kramer considera que en este método las medias marginales dan un peso más adecuado (proporcional) a las observaciones originales que en el de Medias Ponderadas.

Las Sumas de Cuadrados debido a los factores α y β se calculan de manera semejante al método anterior, de tal forma que:

$$\hat{\sigma}^{2\,\prime}_{\alpha} = \sum_i \omega'_i \bar{X}'^{2}_{i.} - \left(\sum_i \omega'_i \bar{X}'_{i.} \right)^2 / \sum_i \omega'_i \; / \; a$$

$$\hat{\sigma}^{2\,\prime}_{\beta} = \sum_j \upsilon'_j \bar{X}'^{2}_{.j} - \left(\sum_j \upsilon'_j \bar{X}'_{.j} \right)^2 / \sum_j \upsilon'_j \; / \; b$$

donde ω'_i , υ'_j están dadas por

$$\omega'_i = n^2_{i.} / \left(\sum_j n^2_{.j} / n_{ij} \right) \quad ; \quad \upsilon'_j = n^2_{.j} / \left(\sum_i n^2_{i.} / n_{ij} \right)$$

De nuevo puede demostrarse que $a\hat{\sigma}^{2\,\prime}_{\alpha}/\sigma^2$ y $b\hat{\sigma}^{2\,\prime}_{\beta}/\sigma^2$ se distribuyen como una χ^2 con a-1 y b-1 grados de libertad respectivamente y que son independientes de SCE. En consecuencia, pueden realizarse pruebas F sobre igualdad de efectos.

Kramer indica que este nuevo método puede proveer pruebas estadísticas más potentes que el método anterior y adicionalmente opina que este método es adecuado cuando a y b son grandes y cuando es aceptable suponer que no existe interacción.

El autor propone en su artículo una forma para comparar los dos métodos anteriores que consiste en calcular la esperanza de las sumas de cuadrados respectivas.

Como resultado de la comparación concluye que si se cumple, que;

$$E(\hat{\sigma}_{\alpha}^{2\prime}) > E(\hat{\sigma}_{\alpha}^{2}) \quad y \quad E(\sigma_{\beta}^{2\prime}) > E(\hat{\sigma}_{\beta}^{2}) \qquad (IV.1.2)$$

cuando la hipótesis nula es falsa entonces, el nuevo método es más potente que el de Medias Ponderadas original. Como en general estas sumas de cuadrados no se pueden comparar numéricamente ya que dependen de los valores de $\alpha_1,\alpha_2,\ldots,\alpha_a$ y $\beta_1,\beta_2,\ldots,\beta_b$ respectivamente, el autor propone una aproximación que reduce el criterio a elegir el nuevo método si:

$$\sum_i \omega_i' - \sum_i \omega_i'^2 / \sum_i \omega_i' > \sum_i \omega_i - \sum_i \omega_i^2 / \sum_i \omega_i$$

$$\sum_j \upsilon_j' - \sum_j \upsilon_j'^2 / \sum_j \upsilon_j' > \sum_j \upsilon_j - \sum_j \upsilon_j^2 / \sum_j \upsilon_j \qquad \text{(IV.1.3)}$$

Así Kramer subraya que el Método Modificado de Medias Ponderadas - es mejor que el Método de Medias Ponderadas cuando no existe interacción y se satisfacen las desigualdades (IV.1.3).

Debe advertirse que en el artículo de Kramer no se especifican las hipótesis que realmente se prueban con los diferentes métodos, y - la optimalidad sólo está en función de las estimaciones de las medias marginales.

IV.2 METODOS DE SUMAS DE CUADRADOS ADITIVOS Y DE MINIMOS CUADRADOS EN EL ANALISIS DE VARIANZA CON DATOS DESBALANCEADOS

El análisis de varianza con datos desbalanceados puede efectuarse a través de varios métodos los cuales, según Goslee y Lucas (1965) se clasifican en dos grupos:

I Los métodos de Sumas de Cuadrados Aditivos

II Los métodos de Mínimos Cuadrados

Los métodos de Sumas de Cuadrados Aditivos producen sumas de cuadrados para efectos principales e interacciones que descomponen la suma de cuadrados que representa la variación entre subclases. Estas sumas de cuadrados no tienen en general, una distribución χ^2 y por lo tanto el cociente de cuadrados medios no tiene una distribución F.

Los métodos de Mínimos Cuadrados por su parte, definen pruebas de hipótesis que tienen una distribución F pero, generalmente, las sumas de cuadrados calculadas no son aditivas, esto es, no constituyen una partición de la suma de cuadrados entre subclases.

En su artículo, Goslee y Lucas tratan las restricciones, las fórmulas de cálculo, el nivel de significancia y la potencia de dos Métodos de Mínimos Cuadrados: el Método de Medias Ponderadas y el Modificado de Medias Ponderadas, y dos Métodos de Sumas de -- Cuadrados Aditivos: el Método de Medias no Ponderadas y el del Número Esperado de Celdas.

Para tal efecto consideran el modelo de efectos fijos con dos -- criterios de clasificación con interacción:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

con

$$e_{ijk} \sim N(0, \sigma^2) \quad \text{e independientes.}$$

$$i = \overline{1, a} \qquad j = \overline{1, b} \qquad k = \overline{1, n_{ij}}$$

En este modelo se involucran las siguientes restricciones en los efectos principales y de interacción

$$\sum_i \omega_i \alpha_i = 0 = \sum_j \upsilon_j \beta_j$$

donde ω_i y υ_j son pesos con la siguiente propiedad:

$$\sum_i \omega_i = 1 = \sum_j \upsilon_j \qquad \omega_i \ , \ \upsilon_j > 0$$

$$\sum_i \omega_i \gamma_{ij} = 0 = \sum_j \upsilon_j \gamma_{ij}$$

Además introducen la siguiente notación:

$$n = \frac{n..}{ab} \quad , \quad p_{ij} = \frac{n_{i.}\, n_{.j}}{n..} \quad , \quad h_{i.} = ab \left(\sum_j \frac{1}{n_{ij}} \right)^{-1}$$

$$h_{.j} = ab \left(\sum_i \frac{1}{n_{ij}} \right)^{-1} \quad , \quad h.. = ab \left(\sum_i \sum_j \frac{1}{n_{ij}} \right)^{-1}$$

Las medias por celda, columna y fila y sus respectivas varianzas - se presentan en la Tabla (4.1). La Suma de Cuadrados del error

$$SCE = \sum_i \sum_j \sum_k \left(\gamma_{ijk} - \bar{\gamma}_{ij.} \right)^2$$

es la misma para todos los métodos.

Las sumas de cuadrados por filas, columnas e interacción para los diferentes métodos se presentan en la Tabla (4.2).

Los autores especifican que los pesos se definen en cada método de acuerdo a la suposición que se hace al definir las medias y las - restricciones. Los pesos usados en las restricciones y para defi-- nir las medias marginales se determinan bajo las suposiciones de

interacción y la importancia relativa de las categorías que definen cada clase en la población. Los métodos de medias ponderadas y medias no ponderadas utilizan los mismos pesos para cada una de las medias por celda y son apropiados cuando es válido suponer esta igualdad de pesos en modelos con o sin interacción. Si no -- existe interacción, Goslee y Lucas, afirman que el método adecuado es el de ajuste de constantes (sección IV.1).

Agregan los autores que los métodos de Medias Ponderadas Modificado y el de Número Esperado de Subclases son adecuados cuando es - correcto suponer proporcionalidad en los pesos. Por lo tanto una elección adecuada de las ponderaciones conducirá a métodos efi--- cientes para el análisis de varianza.

Goslee y Lucas comparan los métodos de Sumas de Cuadrados Aditivos y los métodos de Mínimos Cuadrados de acuerdo al nivel de significancia y a la potencia de las pruebas.

En los métodos de Mínimos Cuadrados no existe mayor dificultad - ya que tienen criterios de prueba exactos. Mientras que en los - Métodos de Sumas de Cuadrados Aditivos existe una alteración en - el nivel de significancia debido a que los cuadrados medios co--- rrespondientes a los efectos no siempre se distribuyen conforme a una χ^2.

Goslee ha investigado en un trabajo previo (1956), el nivel de - significancia del método de medias no ponderadas através de un - método exacto y dos métodos aproximados.

Así de uno de sus métodos aproximados se obtine que los grados - de libertad para el efecto entre filas deberían ser

$$f_1' = \frac{r - 1}{1 + \frac{(r-2)\, c^2}{r-1}}$$

donde

r es el número de celdas

c es el coeficiente de variación de las varianzas de las medias de las celdas

$$c^2 = \sum_i (\sigma_i^2 - \sigma_\cdot^2)^2 / r (\sigma_\cdot^2)^2$$

$$\sigma_\cdot^2 = \sum_i \sigma_i^2 / r \qquad y \qquad \sigma_i^2 = \frac{\sigma^2}{a} \sum_j \frac{1}{n_{ij}}$$

Por lo que el cociente de cuadrados medios tiene una distribu--- ción F con f_1' y $f_2 = n.. - ab$ grados de libertad.

Respecto a la potencia de la prueba, de nuevo, los métodos de Mínimos Cuadrados no presentan problemas, ya que ésta puede determinarse por medio de tablas.

En cambio, para los métodos de Sumas de Cuadrados Aditivos es necesario realizar aproximaciones de distribuciones para estudiar la potencia.

En el artículo de Goslee y Lucas se proponen tres aproximaciones para el método de medias no ponderadas y se presentan cuadros comparativos del valor de la potencia de la prueba calculado por los diferentes métodos para diferentes patrones de datos. En base a dicha comparación Goslee y Lucas opinan que es difícil concluir en forma definitiva acerca de la potencia de los diferentes métodos ya que ésta es en general diferente para cada combinación de efectos y patrones de datos. Esto es, ninguno de los métodos analizados parece resultar uniformemente más potente.

TABLA (4.1)

	NUMERO ESPERADO DE CELDAS	MEDIAS PONDERADAS MODIFICADO	MEDIAS PONDERADAS	MEDIAS NO PONDERADAS
$\bar{Y}_{ij}$	$\sum_{k=1} Y_{ijk}/n_{ij}$	$\sum_{k=1} Y_{ijk}/n_{ij}$	$\sum_{k=1} Y_{ijk}/n_{ij}$	$\sum_{k=1} Y_{ijk}/n_{ij}$
σ^2_{ij}	$\frac{\sigma^2}{n_{ij}}$	$\frac{\sigma^2}{n_{ij}}$	$\frac{\sigma^2}{n_{ij}}$	$\frac{\sigma^2}{n_{ij}}$
$\bar{Y}_{i.}$	$\sum_j \frac{n_{.j}}{n_{..}} \bar{Y}_{ij}$	$\sum_j \frac{n_{.j}}{n_{..}} \bar{Y}_{ij}$	$\sum_j \frac{\bar{Y}_{ij}}{b}$	$\sum_j \frac{\bar{Y}_{ij}}{b}$
σ_i^2	$\sum_j (\frac{n_{.j}}{n_{..}})^2 \frac{\sigma^2}{n_{ij}}$	$\sum_j (\frac{n_{.j}}{n_{..}})^2 \frac{\sigma^2}{n_{ij}}$	$\frac{1}{b^2} \sum_j \frac{\sigma^2}{n_{ij}}$	$\frac{1}{b^2} \sum_j \frac{\sigma^2}{n_{ij}}$
$\bar{Y}_{.j}$	$\sum_i \frac{n_{i.}}{n_{..}} \bar{Y}_{ij}$	$\sum_i \frac{n_{i.}}{n_{..}} \bar{Y}_{ij}$	$\sum_i \frac{\bar{Y}_{ij}}{a}$	$\sum_i \frac{\bar{Y}_{ij}}{a}$
σ^2_j	$\sum_i (\frac{n_{i.}}{n_{..}})^2 \frac{\sigma^2}{n_{ij}}$	$\sum_i (\frac{n_{i.}}{n_{..}})^2 \frac{\sigma^2}{n_{ij}}$	$\frac{1}{a^2} \sum_i \frac{\sigma^2}{n_{ij}}$	$\frac{1}{a^2} \sum \frac{\sigma^2}{n_{ij}}$

TABLA (4.2)

	NUMERO ESPERADO DE CELDAS	MEDIAS NO PONDERADAS
$SC\alpha$	$\sum_i n_{i.} (\bar{Y}_{i.} - \bar{Y}_{..})^2$	$bh_{..} \sum_i (\bar{Y}_{i.} - \bar{Y}_{..})^2$
$SC\beta$	$\sum_j n_{.j} (\bar{Y}_{.j} - \bar{Y}_{..})^2$	$ah_{..} \sum_j (\bar{Y}_{.j} - \bar{Y}_{..})^2$
$SC\gamma$	$\sum_i \sum_j \frac{n_{i.} n_{.j}}{n_{..}} (\bar{Y}_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2$	$h_{..} \sum_i \sum_j (\bar{Y}_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2$
	MEDIAS PONDERADAS MODIFICADO	**MEDIAS PONDERADAS**
$SC\alpha$	$n^2_{.} \Sigma \left(\Sigma \frac{n_{.j}}{n_{ij}} \right)^{-1} \left(\bar{Y}_{i.} - \Sigma_i \left(\Sigma_j \frac{n^2_{.j}}{n_{ij}} \right)^{-1} \bar{Y}_{i.} / \Sigma_i \left(\Sigma_j \frac{n^2_{.j}}{n_{ij}} \right)^{-1} \right)^2$	$b \sum_i h_{i.} \left(\bar{Y}_{i.} - \sum_i h_{i.} \bar{Y}_{i.} / h_{..} \right)^2$
$SC\beta$	$n^2_{.} \Sigma_j \left(\Sigma_i \frac{n_{i.}}{n_{ij}} \right)^{-1} \left(\bar{Y}_{.j} - \Sigma_j \left(\Sigma_i \frac{n^2_{i.}}{n_{ij}} \right)^{-1} \bar{Y}_{.j} / \Sigma_j \left(\Sigma_i \frac{n^2_{i.}}{n_{ij}} \right)^{-1} \right)^2$	$a \sum_j h_{.j} \left(\bar{Y}_{.j} - \sum_j h_{.j} \bar{Y}_{.j} / h_{..} \right)^2$
$SC\gamma$	$\sum_i \sum_j n_{ij} \bar{Y}_{ij.} - \sum_i n_{i.} \bar{Y}_{i..} - r' C^{-1} r$ (ver sección III.2)	

IV.3 EL ANALISIS POR MINIMOS CUADRADOS DE DATOS EXPERIMENTALES

En un trabajo de Overall y Spiegel (1969), parten del hecho bien establecido en la literatura, de que los modelos de diseño de experimentos guardan una cercana relación con los de regresión lineal múltiple. Más aún, en general los modelos de diseño experimental pueden ser expresados como modelos de regresión donde pueden analizarse, de diversas formas los efectos de interés.

Cuando la estructura de las observaciones de que se dispone es balanceada, la ortogonalidad de los efectos implica que el análisis mediante los modelos de regresión es único y coincide con el que se obtiene del análisis de varianza usual. Sin embargo, en los de desbalanceo el análisis por regresión puede efectuarse con diversos procedimientos que usualmente producen resultados distintos que naturalmente, pueden conducir a conclusiones erróneas si no se interpretan adecuadamente.

Por esta razón Overall y Spiegel, en su artículo, proponen una técnica para el análisis de modelos de diseño desbalanceados, a través de regresión, que consideran adecuada y la comparan con otros procedimientos alternativos.

Los autores afirman que los datos desbalanceados producen no ortogonalidad entre los criterios de clasificación y en consecuencia es difícil determinar que factor está influyendo la variable de respuesta bajo prueba y en qué medida. Por tal razón, agregan Overall y Spiegel, pueden usarse los métodos de mínimos cuadrados para estimar los efectos de cada variable ajustados por las relaciones con otras variables de clasificación. No obstante debe de tenerse presente que únicamente bajo circunstancias muy restrictivas los resultados son equivalentes a los del análisis de varianza convencional y, además, las pruebas de hipótesis no son las mismas que se efectúan en los diseños experimentales balanceados usuales.

Overall y Spiegel consideran de interés tres métodos, que llaman de mínimos cuadrados, para el análisis de datos experimentales. Estos producen resultados idénticos, como ya se indicó, al aplicarse a problemas con frecuencias iguales de las observaciones en las celdas, pero conducen a resultados diferentes en otros casos. Los métodos son :

I) Mínimos Cuadrados Completos o Análisis General de Modelos Lineales.

II) Diseño Experimental

III) Ordenación Apriori.

Todos estos métodos son correctos, pero cada uno de ellos prueba un conjunto de hipótesis distinto y las diferencias no son distinguibles en las Tablas de Análisis de Varianza correspondientes. - Los autores, en su artículo, presentan una descripción breve del enfoque por regresión múltiple para el análisis de datos experimentales de tal forma que los efectos principales y las interacciones quedan definidas por variables independientes (indicadoras) en un modelo de regresión. Hay muchas maneras mediante las cuales esas variables independientes pueden ser definidas. Overall y Spiegel proponen generar una matriz de diseño (de variables indicadoras) que impongan las restricciones usuales del análisis de varianza de tal forma que se pueden obtener directamente las estimaciones de los parámetros y las sumas de cuadrados requeridas para los métodos propuestos.

Supóngase el modelo con dos criterios de clasificación con interacción.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk} \qquad \begin{matrix} i = \overline{1,a} \\ j = \overline{1,b} \\ k = \overline{1,n_{ij}} \end{matrix}$$

Con

μ la media general

α_i las desviaciones de las medias de hilera con respecto a la media general

β_j las desviaciones de las medias de columna con respecto a la media general

γ_{ij} son desviaciones de las medias de cada celda con relación a los efectos de hilera y columna.

e_{ijk} es el error aleatorio con distribución normal, media cero, varianza σ^2 e independiente.

Para obtener una solución por regresión en la cual los coeficientes estimados sean a su vez los estimadores de los parámetros involucrados en el modelo de diseño experimental, es necesario imponer las restricciones adecuadas. Un conjunto de restricciones acostumbrado es el siguiente:

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$$

La matriz de diseño que incluye estas restricciones se forma, de acuerdo a los autores, incluyendo inicialmente, a-1 columnas que representan los primeros a-1 niveles de α y b-1 niveles de β. Para cada observación, los valores de las columnas asociadas con los efectos principales de α y β se determinan de tal forma que si la observación pertenece a una de las primeras a-1 categorías de α, se le asigna un uno en la columna correspondiente y

cero en las demás columnas asociadas con α . Si el individuo pertenece a la última categoría de α se le asigna el valor menos uno (-1) en todas las a-1 columnas asociadas con el efecto principal α. De modo similar se procede para el efecto β . Las entradas de las columnas de la matriz de diseño correspondientes a la interacción γ son (a-1)(b-1) y se obtienen como producto de las entradas en las columnas correspondientes de los efectos principales. De esta manera, la matriz de diseño resulta de rango completo y pueden obtenerse los estimadores de los parámetros del modelo así como las sumas de cuadrados debidas a los subconjuntos de variables independientes de interés. Este procedimiento para transformar el modelo de diseño en uno de regresión es común a los tres métodos de análisis examinados por Overall y Spiegel. Ahora bien, las diferencias en los resultados se deben, como ya se indicó, al procedimiento particular de análisis en el modelo de regresión.

Cabe recordar que en los modelos de regresión donde se tiene una variable respuesta Y en términos de k variables independientes $(X_1, X_2, \ldots, X_k)$ se utilizan determinados procedimientos estadísticos para seleccionar las variables más significativas. Algunos de los métodos más frecuentes en la literatura son: [Draper (1966)]

i) El de selección hacia adelante (forward)

ii) El de selección hacia atrás (backward)

iii) El de selección paso a paso (stepwise)

Estos procedimientos pueden utilizarse en el modelo de diseño - transformado, pero debe subrayarse que en este caso no es de interés estudiar la contribución particular de una variable, sino del conjunto de éstas que determinan los efectos de un factor. - Los métodos que sugieren Overall y Spiegel se basan en esta observación y las diferencias se presentan en la forma de evaluar las contribuciones de los diversos factores. El Método 1 involucra la estimación de efectos de cada factor ajustado por todos - los demás incluídos en el modelo. Puede advertirse la semejanza de este método con el método hacia atrás.

El Método 2, involucra la estimación de efectos principales ignorando interacciones y después la estimación de interacciones ajustado por efectos principales. En este caso, se puede decir, - que se trabaja con dos modelos: uno que contiene únicamente efectos principales y otro que incluye además el término de interacción.

El Método 3, considera un ordenamiento inicial de los efectos - en el modelo y realiza la estimación de los efectos ajustados - por los efectos anteriores. Puede pensarse en este método como similar al de hacia adelante.

Overall y Spiegel opinan que existen pocas bases para decir que - alguno de los tres métodos resulta en general, en el análisis de varianza, el más adecuado. Este hecho se debe principalmente a - que cada uno corresponde a una estrategia distinta. Los autores presentan un ejemplo tomado de Linquist (1953) en el cual las celdas tienen frecuencias proporcionales y en consecuencia el problema puede ser manejado por el análisis de varianza convencional. - Del estudio, Overall y Spiegel opinan que el Método 2 parece ser la generalización más apropiada del análisis de varianza usual - del diseño experimental. Esto se debe a que cuando se tienen frecuencias proporcionales, los efectos principales son ortogonales entre sí, aunque las frecuencias desiguales producen correlaciones entre las interacciones y los efectos principales.

Finalmente Overall y Spiegel exponen algunas características distintivas de los diferentes métodos.

En lo que se refiere al método 1, los autores opinan, que es adecuado si el problema se concibe en términos de un modelo de regresión lineal múltiple. Adicionalmente este método puede utilizarse para determinar si los factores tienen efectos principales persistentes, que no son explicados en su totalidad, mediante efectos de interacción.

En cuanto al método 2, Overall y Spiegel, consideran que es apropiado cuando se espera encontrar variación significativa en términos de efectos principales aditivos.

El método 3 tiene una ventaja pronunciada al minimizar la posibiliad de que efectos significativos se cancelen unos a otros. Al trabajar con este método, las sumas de cuadrados de las componentes, siempre al sumarse producen la suma de cuadrados del total.

En los tres métodos debe tenerse cuidado de no interpretar los efectos principales significativos de la misma manera que se interpretan en un modelo balanceado.

Por último, los autores consideran que la más directa generalización del análisis de varianza convencional para los diseños experimentales lo proporciona el método 2. Sin embargo, afirman que dependiendo de los intereses del investigador pueden utilizarse el método 1 o bien el 3. Asimismo sugieren, dadas las diferencias existentes entre los datos, que siempre se especifique qué método se utilizó en el análisis de tal forma que pueda realizarse una interpretación correcta de los resultados.

IV.4 ANALISIS DE VARIANZA NO ORTOGONAL USANDO MEJORAMIENTO ITERATIVO Y RESIDUALES BALANCEADOS

Hemmerle (1974) presenta en su trabajo un método para realizar el análisis de varianza no ortogonal (desbalanceado) usando las medias por celda. Una de las ventajas del procedimiento es que evita el almacenamiento de gran cantidad de información, al no requerir del cálculo de la multiplicación de la matriz de diseño por su transpuesta ($X_0'X_0$), o bien de una transformación ortogonal para X_0. El método es iterativo y convergente. Utiliza los estimadores y los residuales del análisis de varianza balanceado para resolver las ecuaciones normales y para realizar pruebas de hipótesis, con la propiedad, de que minimiza las iteraciones en las pruebas de hipótesis en caso de tener factores o interacciones no significativas.

Supóngase el modelo de efectos fijos con dos criterios de clasificación e interacción

$$Y_{ijk} = \mu_{ij} + e_{ijk} \qquad \begin{matrix} i = 1,\ldots,a \\ j = 1,\ldots,b \\ k = 1,\ldots,n_{ij} \end{matrix} \qquad (IV.4.1)$$

Con

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

$$E(e_{ijk}) = 0 \qquad E(e_{i'j'k'} \cdot e_{ijk}) = 0$$

para $(i',j',k') \neq (i, j, k)$

$$E(e^2_{ijk}) = \sigma^2$$

O bien en forma matricial

$$Y = X_0\beta + e \qquad . \quad . \quad . \qquad (IV.4.2)$$

Donde

Y es el vector de observaciones de dimensión N

X_0 es la matriz de diseño de orden N x p

β es un vector de dimensión p de parámetros del modelo.

Si el modelo (IV.4.2) se reparametriza utilizando el conjunto de restricciones usuales:

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$$

y así obtener una matriz X_0 de rango completo (p).

Entonces la solución para las ecuaciones normales en (IV.4.2) es

$$\hat{\beta} = (X_0' X_0)^{-1} X_0' Y.$$

Si

$$\bar{Y}_{ij.} = \frac{1}{n_{ij}} \sum_k Y_{ijk}$$

reescribiendo el modelo (IV.4.2) basándose en las medias de las - celdas se tiene que

$$\bar{Y}_{ij.} = \mu_{ij} + \delta_{ij} \quad . \quad . \quad . \quad (IV.4.3)$$

donde

$$E(\delta_{ij}) = 0, \qquad E(\delta_{i'j'}, \delta_{ij}) = 0$$

$$E(\delta_{ij}^2) = \sigma^2 / n_{ij}$$

La representación matricial de las ecuaciones correspondientes a los mínimos cuadrados ponderados para el modelo (IV.4.3) es la siguiente:

$$X' D X \hat{\beta} = X' D \bar{Y} \qquad (IV.4.4)$$

donde

X es una matriz n x p de diseño que se obtendría para el modelo (IV.4.1) si hubiera una sola observación por celda.

D es una matriz diagonal que contiene las frecuencias de las celdas. Esto es, sus elementos son: $d_1 = n_{11}$, $d_2 = n_{12}$, ..., $d_n = n_{ab}$ $n_{ij} \neq 0$.

β es el vector de dimensión p de parámetros -- desconocidos contenidos en μ_{ij} y

$\bar{Y}$ es el vector de dimensión n de las medias de las celdas.

Puede verificarse que

$$X_0' X_0 = X'DX$$

y

$$X_0' Y = X'D\bar{Y}$$

en consecuencia

$$\hat{\beta} = (X_0' X_0)^{-1} X_0' Y = (X' D X)^{-1} X' D \bar{Y} .$$

donde la suma de cuadrados debido al ajuste del modelo completo es

$$S = \hat{\beta}' X_0' Y = \hat{\beta}' X' D \bar{Y}$$

y la suma de cuadrados de los residuales, R_s es igual a

$$R_s = Y' Y - \hat{\beta}' X_0 Y = Y' Y - \hat{\beta}' X' D \bar{Y}$$

Para el caso especial en el cual X es cuadrada y no singular, se puede premultiplicar ambos lados de (IV.4.3) por $(X')^{-1}$ y después por D^{-1} para obtener

$$X \hat{\beta} = \bar{Y}$$

Premultiplicando ambos lados por X' se obtiene

$$X' X \hat{\beta} = X' \bar{Y}$$

En este caso, los estimadores $\hat{\beta}$ para el modelo no ortogonal son los mismos que aquéllos obtenidos de (IV.4.3) usando procedimientos balanceados. Sin embargo cuando estas circunstancias no se presentan, Hemmerle propone un método iterativo para resolver el sistema de ecuaciones (IV.4.4), para ello parte del supuesto de que

$$A_0 \approx (X' D X)^{-1}$$

es una inversa aproximada con la propiedad de que

$$\hat{\beta}_0 = A_0 X' D \bar{Y}$$
$$\hat{\beta}_1 = (I + E_0) A_0 X' D \bar{Y},$$
$$\vdots$$
$$\hat{\beta}_k = (I + E_0 + E_0^2 + \ldots + E_0^k) A_0 X' D \bar{Y}$$

donde $\hat{\beta}_i$ es la i-ésima aproximación al valor de $\hat{\beta}$ y

$$Eo = (I - Ao\ X'\ D\ X)\ ;$$

además

$$I + Eo + Eo^{2} + \ldots + Eo^{k} + \ldots = (I - Eo)^{-1}$$

siempre y cuando las raíces características de Eo, λ_k, sean menores a uno. Hemmerle demuestra que $\lambda_k < 1\ \forall\ k = 1,\ldots,p$ por consiguiente

$$\hat{\beta}_k = (I - Eo)^{-1}\ Ao\ X'\ D\ \bar{Y}$$
$$= (X'\ D\ X)^{-1} X'\ D\ \bar{Y} = \hat{\beta}$$

Por otra parte, el autor propone como inversa aproximada inicial- $Ao = (1/c)\ (X'X)^{-1}$ a fin de que se faciliten los cálculos. - Para ello, y con el propósito de que la serie converja se determina el valor de c como

$$c = \max \{ d_i \}$$

Hemmerle sugiere una técnica para calcular, en forma sencilla, el valor del(k+1) - ésimo elemento de la serie, a partir del k-ésimo. Este procedimiento utiliza algunas formas conocidas del análisis de varianza balanceado.

$$\hat{\beta}_{k+1} = \hat{\beta}_0 + \hat{\beta}_k - E\,[(\,D\,/\,c\,)\,V_k\,]$$

$$\hat{\beta}_0 = E\,[(\,D\,/\,c\,)\,\bar{Y}\,]$$

donde

$$V_k = V_0 + [\,I - (\,D/c\,)]\,V_{k-1} + R\,[\,(\,D/c\,)\,V_{k-1}\,] = X\hat{\beta}_k$$

$$V_0 = (\,D/c\,)\,\bar{Y} - R\,[(\,D/c)\,\bar{Y}\,]$$

$$R = [\,I - X\,(X'X)^{-1}X'\,] \qquad \ldots \qquad (IV.4.5)$$

$$E = (\,X'X\,)^{-1}\,X'$$

Para realizar pruebas de hipótesis que involucren a los estimadores de los parámetros, Hemmerle propone otro método iterativo, que consiste en calcular la suma de cuadrados debido al modelo restringido por las hipótesis de interés:

$$S_k = \bar{Y}'\,D\,V_k \qquad \text{en la k-ésima iteración}$$

con

$$S_0 = \bar{Y}'\,D\,V_0 \qquad \ldots \qquad (IV.4.6)$$

en donde

V_k se obtiene de (IV.4.5)

Por último, el autor en su artículo presenta el siguiente cuadro que ilustra el procedimiento para el caso del modelo descrito en (IV.4.1)

HIPOTESIS	MODELO RESTRINGIDO	FORMA DE LOS RESIDUALES BALANCEADOS	ESTADISTICA F
$Ho : \alpha_i = 0 \quad \forall\, i$	$Y_{ijk} = \mu + \beta_j + \gamma_{ij} + e_{ijk}$	$V_{i.} - V..$	$\dfrac{(\tilde{Y}'D\tilde{Y} - S^{\alpha}) / a - 1}{Y'Y - \tilde{Y}'D\tilde{Y} / N - ab}$
$Ho : \beta_j = 0 \quad \forall\, j$	$Y_{ijk} = \mu + \alpha_i + \gamma_{ij} + e_{ijk}$	$V_{.j} - V..$	$\dfrac{(\tilde{Y}'D\tilde{Y} - S^{\beta}) / b - 1}{Y'Y - \tilde{Y}'D\tilde{Y} / N - ab}$
$Ho : (\gamma_{ij}) = 0 \quad \forall\, (i,j)$	$Y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$	$V_{ij} - V_{i.} - V_{.j} + V..$	$\dfrac{(\tilde{Y}'D\tilde{Y} - S^{\gamma}) / (a-1)(b-1)}{Y'Y - \tilde{Y}'D\tilde{Y} / N - ab}$

Donde S^{α}, S^{β} y S^{γ} se obtienen de (IV.4.) al imponer las restrcciones $\alpha_i = 0 \ \forall\, i$, $\beta_j = 0 \ \forall\, j$, $\gamma_{ij} = 0 \ \forall\, i,j$ al modelo completo.

IV.5 EL ANALISIS DE RANGO COMPLETO E INCOMPLETO PARA MODELOS LINEALES

En un artículo de Speed y Hocking (1975) se exponen algunos ejemplos para comparar el análisis de varianza de rango incompleto - con el de rango completo o análisis de medias. Particularmente - los autores consideran que el análisis de rango completo elimina la confusión cuando se realizan pruebas de hipótesis que involu-- cran datos desbalanceados.

Supóngase el modelo

$$Y = X\beta + e \qquad \ldots \qquad (IV.5.1)$$

con

- Y un vector de $n \times 1$ observaciones
- e un vector $n \times 1$ de errores, además $e \sim N(0, \sigma^2 I)$
- β es un vector $p \times 1$ de parámetros desconocidos
- X es una matriz de diseño de orden $n \times p$ y rango $q < p$

El propósito fundamental del artículo no es realizar una aporta-- ción teórica matemática para el análisis del modelo (IV.5.1), - sino establecer una mediación entre la teoría y la práctica. Espe cíficamente, los autores intentan aclarar algunas confusiones -

relacionadas con las pruebas de hipótesis así como simplificar - las ideas acerca del análisis de datos desbalanceados.

Para realizar la discusión Speed y Hocking introducen la siguiente notación alternativa al modelo (IV.5.1) de la siguiente manera:

$$Y = W\mu + e \qquad \ldots \qquad (IV.5.2)$$

$$\text{s.a.} \quad G\mu = 0$$

Donde

Y , e son las mismas que en (IV.5.1)

μ es el vector de medias cuya longitud es igual al número de poblaciones muestreadas

W es una matriz de unos y ceros tal que cada columna tiene tantos unos como observaciones en la población correspondiente

G expresa la relación entre las medias que - depende de la situación particular considerada. G es de orden r x m con rango r.

Puede observarse que el modelo (IV.5.1) es una sobreparametrización del Modelo (IV.5.2) lo cual conduce a conceptos tales como - funciones no estimables e hipótesis no probables que causan confusión particularmente en los modelos desbalanceados. Cuando se - realizan pruebas de hipótesis el desarrollo teórico proporciona -

expresiones explícitas para hipótesis de la forma $K\beta = n$. En la práctica afirman los autores, las hipótesis que así se prueban son poco claras lo que conduce a adoptar algunas medidas computacionales que prueban hipótesis que no necesariamente son las de interés.

Para realizar pruebas de hipótesis de la forma $\Lambda\beta = n$ en el modelo (IV.5.1), la prueba estadística es

$$[(R_H^2 - R_0^2) / R_0^2] (q / s) \quad . . . \quad (IV.5.3)$$

donde q y s son el rango de X y Λ respectivamente y

$$R_0^2 = \min_{\beta} (Y - X\beta)' (Y - X\beta);$$

$$R_H^2 = \min_{\beta} (Y - X\beta)' (Y - X\beta)$$

$$s.a \qquad \Lambda\beta = n$$

Mientras que para probar hipótesis de la forma $\Omega\mu = \xi$ en el modelo (IV.5.2), es necesario elegir Ω tal que $\begin{bmatrix} \Omega \\ G \end{bmatrix}$ tenga rango completo. Speed y Hocking estudian la falta de claridad en el análisis del modelo (IV.5.1) aún en los modelos con un criterio de clasificación. Consideran las dos representaciones del modelo correspondiente:

$$Y_{ij} = \mu + \alpha_i + e_{ij} \qquad i = \overline{1,a} \; ; \; j = \overline{1,n_i} \qquad (IV.5.4)$$

$$Y_{ij} = \mu_i + e_{ij} \qquad i = \overline{1,a} \; ; \; j = \overline{1,n_i} \qquad (IV.5.5)$$

Con μ_i estimable si y sólo si $n_i > 0$. El modelo sobreparametrizado (IV.5.4) se obtiene de establecer $\mu_i = \mu + \alpha_i$ en el modelo (IV.5.5) . Como ya se sabe la elección de las condiciones no estimables no influye el análisis, sin embargo, la elección de alguna condición puede conducir a equívocos. Así de acuerdo al artículo de Speed y Hocking, la cantidad $(\mu_1 + \mu_2) / 2$ es estimable, si se reparametriza imponiendo la condición $\alpha_1 = 0$, esta cantidad se describe como $\mu + \alpha_2 / 2$, pero si se impone la condición $\alpha_1 + \alpha_2 = 0$ se describe como μ . Las dos cantidades describen a $(\mu_1 + \mu_2) / 2$, la elección de la condición no estimable asigna un significado a μ y α_i y este si difiere bajo las dos elecciones.

En el caso de los modelos con dos criterios de clasificación los autores también consideran los dos tipos de representaciones:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ij} \qquad \ldots \qquad (IV.5.6)$$

$$Y_{ijk} = \mu_{ij} + e_{ij} \qquad i = \overline{1,a} \quad j = \overline{1,b} \qquad \ldots \qquad (IV.5.7)$$

sin restricción en μ_{ij}

Cuando se analiza el caso balanceado es usual considerar hipóte-- sis sobre los efectos principales α y β y sobre interacción. Los autores ilustran los errores a los que se pueden llegar debido a la falta de comprensión en el concepto de imposición de condiciones no estimables en el modelo (IV.5.6).

Speed y Hocking fijan su atención en los cuadrados medios asocia dos al efecto principal α ($CM\alpha$) y encuentran que su valor espera do, $ECM\alpha$ en algunos textos es consignado como:

$$\sigma^2 + \frac{n\,b}{a-1} \sum_i \alpha_i^2 \qquad \ldots \qquad (IV.5.8)$$

y en otros como

$$\sigma^2 + \frac{n\,b}{a-1} \sum_i (\alpha_i - \bar{\alpha}_\cdot + \gamma_{i\cdot} - \bar{\gamma}_{\cdot\cdot})^2 \qquad (IV.5.9)$$

La ecuación (IV.5.8) sugiere que $CM\alpha$, es apropiada para probar i- gualdad de efectos, mientras que la ecuación (IV.5.9) sugiere que la hipótesis que se prueba es la de igualdad de tratamientos in-- fluenciada por el promedio de las interacciones.

Las condiciones usuales impuestas en el modelo son

$$\alpha_\cdot = \beta_\cdot = \gamma_{i\cdot} = \gamma_{\cdot j} = 0 \qquad \forall\ i,j \qquad (IV.5.10)$$

de donde se puede observar que (IV.5.8) es la expresión correcta -

para ECM_α si se reparametriza el modelo imponiendo (IV.5.10). Por otro lado, (IV.5.9) es correcta si no se impone ninguna condición al modelo. Speed y Hocking enfatizan que el hecho de establecer - condiciones no estimables no conduce a errores, es más bien la - técnica computacional adoptada para reducir el modelo (IV.5.6) en un modelo de rango completo lo que ocasiona el cambio en las hipótesis que se prueban. Los autores opinan que es frecuente que se determinen incorrectamente las hipótesis asociadas a los cuadrados medios CM_α, CM_β, CM_γ como:

$$H_\alpha = \alpha_i = 0 \qquad H_\beta = \beta_j = 0 \qquad H_\gamma = \gamma_{ij} = 0$$

$$i=\overline{1,a} \; ; \; j=\overline{1,b} \qquad \text{(IV.5.11)}$$

Estas son afirmaciones incompletas de las hipótesis que se están - considerando y sólo tienen sentido si se asocian con la condición (IV.5.10). En términos del modelo (IV.5.2) las hipótesis completas se representan por:

$$H_\alpha: \mu_{i.} = \mu_{i'.} \qquad H_\beta: \mu_{.j} = \mu_{.j'}$$

$$H_\gamma: \mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} = 0 \qquad \text{(IV.5.12)}$$

$$\forall \; i,i' \; ; \; j, j'$$

En términos del modelo (IV.5.1), si se establece la relación - $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$ a (IV.5.12) y además se reparametriza el modelo imponiendo la condición (IV.5.10) se obtiene el mismo resultado que en (IV.5.11).

Sin embargo, una confusión que a menudo se sucita es debida a que no todas las condiciones no-estimables cuando se imponen en el - modelo (IV.5.6) conducen a la misma equivalencia entre (IV.5.11) y (IV.5.12). Speed y Hocking proponen como una alternativa a - (IV.5.10) las siguientes condiciones no-estimables:

$$\alpha_a = \beta_b = \gamma_{ib} = \gamma_{aj} = 0 \quad . \quad . \quad . \quad \text{(IV.5.13)}$$

Harvey (1960) opinó que el uso de (IV.5.13) conduce a un serio error al producir sumas de cuadrados incorrectas. No obstante, Speed y Hocking afirman que éstas no son incorrectas sino que corresponden a hipótesis diferentes que no son las de interés.

Speed y Hocking exponen otro ejemplo que considera el análisis - del modelo (IV.5.6) desbalanceado. Discuten un ejemplo presentado por Chakravarti (1967) en el cual existe una confusión debido a las hipótesis que se prueban. Presentan una Tabla de Análisis de Varianza con los siguientes datos:

FUENTE	GRADOS DE LIBERTAD	CUADRADOS MEDIOS
Sujetos (ajustados)	2	10,149.22
Observadores (ajustados)	2	254.18
Interacción	4	177.48
Error	13	92.62

Si el modelo se representa de acuerdo al modelo (IV.5.2) una elección usual de la hipótesis que se prueba podría ser (IV.5.12). Sin embargo, esto no es correcto ya que se debe considerar las - frecuencias de las celdas.

Una segunda alternativa que se plantea en el artículo es considerar

$$H_\alpha^* : \sum_j \frac{n_{ij}}{n_{i.}} \mu_{ij} = \sum_j \frac{n_{i'j}}{n_{i'.}} \mu_{i'j} \quad . \quad . \quad . \qquad \text{(IV.5.14)}$$

$$H_\beta^* : \sum_i \frac{n_{ij}}{n_{.j}} \mu_{ij} = \sum_i \frac{n_{ij'}}{n_{.j'}} \mu_{.j'}$$

No obstante estas hipótesis no conducen a los cuadrados medios de la tabla sino que pertenecen a los cuadrados medios llamados debido al tratamiento α no ajustado, y debido al tratamiento β no a-- justado.

Finalmente, de acuerdo a lo establecido por Searle (1971) los - cuadrados medios que se presentan en la tabla corresponden respectivamente a las hipótesis

$$H_\alpha^{**}: \sum_j \left(n_{ij} - \frac{n_{ij}^2}{n_{.j}}\right) \mu_{ij} - \sum_{i' \neq i} \sum_j \frac{n_{ij} n_{i'j}}{n_{.j}} \mu_{i'j} = 0$$

$$\text{(IV.5.15)}$$

$$H_\beta^{**}: \sum_i \left(n_{ij} - \frac{n_{ij}^2}{n_{i.}}\right) \mu_{ij} - \sum_{j' \neq j} \sum_i \frac{n_{ij} n_{ij'}}{n_{i.}} \mu_{ij'} = 0$$

Cuando se trabaja con datos desbalanceados, Speed y Hocking opinan que las hipótesis adecuadas son las establecidas en (IV.5.14) ya que las presentadas en (IV.5.15) son de difícil interpretación.

A través de los ejemplos presentados en este trabajo, los autores comparan el análisis de los modelos de diseño de experimentos vía el modelo (IV.5.1) contra el modelo (IV.5.2) de rango completo.

IV.6 UNA ALTERNATIVA PARA EL ANALISIS DE MODELOS LINEALES QUE CONTIENEN CELDAS VACIAS

Speed, Hocking y Coleman (1980) han propuesto una alternativa para el análisis de modelos lineales que contienen celdas vacías y proporcionan un criterio para seleccionar las hipótesis que se prueban en un caso desbalanceado.

Los autores suponen que el modelo se establece en términos del vector de parámetros μ y que la hipótesis que se pretende probar es de la forma $H\mu = 0$ por lo que la prueba estadística está dada por

$$F_{(q,n-p)} = \frac{SC(H)}{qs^2}$$

donde

- q es el rango de H
- s^2 son los cuadrados medios del residual con n-p grados de libertad del modelo original.
- $SC(H)$ es la suma de cuadrados del numerador

$$SC(H) = (H\hat{\mu})' (HV(\hat{\mu}) H')^{-} H\mu$$

Por lo general, los diversos programas de computadora no especifican la matriz de hipótesis H y SC(H) no se calcula directamente, sino que la cantidad que se utiliza como la suma de cuadrados del numerador se obtiene como una consecuencia de un procedimiento computacional particular.

Con el fin de conocer las hipótesis que realmente se prueban, los autores recomiendan, si no hay celdas vacías, que se utilicen las hipótesis que podrían probarse en el caso balanceado. Si existen celdas vacías, los autores proponen un procedimiento para determinar una hipótesis razonable que sea similar a la hipótesis deseada.

Para tal efecto Speed, Hocking y Coleman utilizan el modelo de medias de rango completo:

$$Y = W\mu + e$$

s. a

$$G\mu = 0$$

donde

μ es el vector de dimensión p de las medias de las celdas de la población que pueden incluirse en el estudio.

W es una matriz de conteo tal que $W'W$ es diagonal y es igual al número de observaciones en la i-ésima población.

G es una matriz r x p de rango r tal que $G\mu = 0$ representa un conjunto de relaciones entre las medias de las celdas conocidas.

e es el error aleatorio con $e \sim N(0, I\sigma^2)$ e independientes.

Como primer paso los autores determinan las restricciones de las medias en la población observada que están implícitas en $G\mu = 0$. Para ello particionan μ en $\mu' = (\mu_m', \mu_o')$ donde μ_m y μ_o corresponden respectivamente a las medias de las celdas vacías y las observadas. En base a lo anterior propocionan la siguiente definición:

> " La restricció $\tilde{G}\mu_o = 0$ se dice que es la restricción efectiva si cuando $G\mu = 0$ implica $\tilde{G}\mu_o = 0$ y $\tilde{G}$ es de rango máximo."

Con esta definición y asumiendo que W_o es la matriz de conteo apropiada, el modelo efectivo es

$$Y = W_o\mu + e$$

$$\text{s. a} \quad \tilde{G}\mu_o = 0$$

Para construir G se particiona G en

$$G = [\, Gm, Go \,],$$

de acuerdo al mismo criterio con el que se particionó μ, se aplican operaciones entre filas con el objeto de reducir G a la forma:

$$\begin{bmatrix} Gmm & Gmo \\ 0 & \tilde{G} \end{bmatrix} \quad . \ . \ . \quad (IV.6.1)$$

Para probar una hipótesis en el modelo efectivo cuyo rechazo implique el rechazo de $H\mu = 0$ en el modelo original, Speed Hocking y Coleman proporcionan la siguiente definición:

" La hipótesis $\tilde{H}\mu_0 = 0$ se dice es la hipótesis efectiva si $\tilde{H}\mu_0 \neq 0 \Rightarrow H\mu \neq 0$ y H es de rango máximo "

De la matriz (IV.6.1) se observa que μ_m está definida en términos de μ_0 como:

$$Gmm\mu_m + Gm_0\mu_0 = 0 \qquad (IV.6.2)$$

Con μ_m definida en forma única si Gmm es no singular.
Si se utiliza la ecuación en (IV.6.2) para eliminar todo o parte - de μ_m de las ecuaciones se obtiene

$$H^*\mu^* = H^*m\,\mu_m^* + Ho\mu_0 = 0$$

Nótese que si Gmm es no-singular, $H^*m = 0$ y $\tilde{H} = H^*o$. En general, aplicando las reducciones en las filas a H^* se puede obtener:

$$\begin{bmatrix} \overset{\blacktriangle}{H}_{mm} & \overset{\blacktriangle}{H}_{mo} \\ 0 & \tilde{H} \end{bmatrix}$$

y queda especificado $\tilde{H}$.

La metodología que Speed, Hocking y Coleman presentan en su artículo tiene la particularidad de que las hipótesis efectivas se basan en hipótesis que se habrían probado si el experimento fuera balanceado.

CONCLUSIONES

CONCLUSIONES

El problema del desbalanceo ha sido objeto de un considerable número de contribuciones en la literatura estadística. Sin embargo, y dado que el análisis parece complicarse en la medida en que aumenta la cantidad de factores bajo consideración en el modelo, prácticamente la totalidad del material que ha sido recopilado para la elaboración de este trabajo sólamente aborda el caso de dos factores. De esta forma, las conclusiones que aquí pueden presentarse son relativas a este tipo de modelos.

La manera teóricamente correcta en la que deben de analizarse los modelos de diseño de experimentos desbalanceados es, de acuerdo a lo expuesto en el apartado II.3, la correspondiente al modelo general. Esto es, se debe de establecer la hipótesis de la forma H: Kb = m y a partir de ella obtener las estadísticas de prueba. Sin embargo, debido a que de esta manera cada problema particular requiere de la elaboración de programas de computación típicamente complicados, se ha optado por utilizar algunos métodos alternativos usualmente incorporados en paquetes estadísticos que facilitan el análisis pero que pueden producir resultados erróneos si no se utilizan adecuadamente, es decir, si no se tiene claridad en las hipótesis que se prueban.

De cualquier manera, a continuación se presenta un resumen de las hipótesis que se prueban con los diferentes métodos expuestos en este trabajo, para lo cual es útil la siguiente tabla:

$$H_1: \bar{\mu}_{i.} = \bar{\mu}_{i'.}$$

$$H_2: \sum_j n_{ij}\mu_{ij} / n_{i.} = \sum_j n_{i'j}\mu_{i'j} / n_{i'.}$$

$$H_3: \sum_j (n_{ij})\mu_{ij} = \sum_{i'} \sum_j \frac{n_{ij} n_{i'j} \mu_{i'j}}{n_{.j}}$$

$$H_4: \mu_{i1} = \mu_{i'1}$$

$$H_5: \bar{\mu}_{.j} = \bar{\mu}_{.j'}$$

$$H_6: \sum_i n_{ij}\mu_{ij} / n_{.j} = \sum_i n_{ij'}\mu_{ij'} / n_{.j'}$$

$$H_7: \sum_i (n_{ij})\mu_{ij} = \sum_{j'} \sum_i \frac{n_{ij} n_{ij'} \mu_{ij'}}{n_{i.}}$$

$$H_8: \mu_{1j} = \mu_{1j'}$$

$$H_9: \mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} = 0$$

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

El método de medias no ponderadas es un método aproximado puesto - que las sumas de cuadrados no tienen una distribución χ^2 y en consecuencia no corresponden al numerador de una prueba de hipótesis lineal. No obstante Goslee y Lucas sugirieron una aproximación para probar las hipótesis H_1 y H_5 . La primera justificación para el método parecía ser su simplicidad en el cálculo; sin embargo con - el avance en las actuales computadoras esta ventaja parece no ser de gran importancia.

El método de número esperado de celdas no es útil para realizar - pruebas de hipótesis ya que, de forma similar que en el anterior, la suma de cuadrados del numerador no tiene una distribución χ^2.

El método de medias ponderadas es sencillo y sirve para probar las hipótesis no ponderadas H_1, H_5 y H_9.

Por otro lado el método modificado de medias ponderadas se puede - utilizar para probar las hipótesis

$$H : \alpha_j + \sum_j \frac{n_{\cdot j} Y_{ij}}{n_{\cdot\cdot}} = \alpha_{j'} + \sum_j \frac{n_{\cdot j} Y_{i'j}}{n_{\cdot\cdot}}$$

$$H : \sum_i \frac{n_{i\cdot} Y_{ij}}{n_{\cdot\cdot}} + \beta_j = \sum_i \frac{n_{i\cdot} Y_{ij'}}{n_{\cdot\cdot}} + \beta_{j'}$$

y H_9

El método de ajuste de constantes es el óptimo para los modelos - con dos criterios de clasificación sin interacción. Proporciona el mismo conjunto de sumas de cuadrados que se obtienen al esta-- blecer las hipótesis

$$Ho : \alpha_i - \alpha_a = 0 \qquad \forall\, i = 1,2,\ldots,a-1$$

$$Ho : \beta_j - \beta_b = 0 \qquad \forall\, j = 1,2,\ldots,b-1$$

en el modelo general.

Los métodos de Overall y Spiegel prueban las siguientes hipótesis:

METODO 1. Mínimos Cuadrados Completos. Este método proporcio na las mismas sumas de cuadrados que el de medias ponderadas y, por tanto, sirve para probar el mismo conjunto de hipóte-- sis.

METODO 2. Del Diseño Experimental. Este método es semejante al método de ajuste de constantes y prueba las hipótesis H_3 y H_7.

METODO 3. Ordenación Apriori. Este es también un caso espe-- cial del método de ajuste de constantes en donde se utiliza - $R(\alpha/\mu)$ y $R(\beta/\mu,\alpha)$ como la suma de cuadrados del efecto princi pal que es útil para probar H_2 y H_7.

Alternativamente se puede utilizar $R(\beta/\mu)$ y $R(\alpha/\mu,\beta)$ para probar H_4 y H_3.

El método iterativo de Hemmerle proporciona aproximaciones de las sumas de cuadrados. Las hipótesis que se prueban dependen de - las condiciones no estimables seleccionadas y del orden en el que se particionen los parámetros. En su trabajo de 1974 las sumas - de cuadrados que se obtienen sirven para probar las hipótesis H_1, H_4 y H_3, ya que el conjunto de condiciones no estimables es

$$\alpha_{\cdot} = \beta_{\cdot} = \gamma_{i\cdot} = \gamma_{\cdot j} = 0 \quad \forall\, i,j.$$

Se puede observar que las hipótesis H_1 y H_3 son fáciles de interpretar, además de que las sumas de cuadrados asociadas a ellas - pueden obtenerse haciendo uso de algún paquete de regresión múltiple. Sin embargo debe de tenerse presente la reparametrización - que se está considerando al imponer la condición $\alpha_{\cdot}=\beta_{\cdot}=\gamma_{i\cdot}=\gamma_{j}=0$. Si en efecto es correcto de acuerdo al modelo, establecer esta - restricción este puede ser un método adecuado y de cálculo sencillo.

BIBLIOGRAFIA

BIBLIOGRAFIA

Bicking, C. A., (1954). Some uses of Statistics in the Planning - of Experiments. Industrial Quality Control.

Chakravarti, I. M. (1967). Handbook of Methods of Applied Statistics New York: John Wiley & Sons, Inc.

Draper, N. R. and Smith H. (1966). Applied Regression Analysis . - John Wiley & Sons, Inc.

Goslee D. G. and Lucas H. L. (1965). Analysis of Variance of - Disproportionate Data When Interaction is Present. Biometrics.

Goslee D. G. (1956). The level of significance and Power of the - Unweighted Means Test. Unpublished Ph. D. Thesis. North Carolina State College, Raleigh, N. C.

Harvey, W. R. (1960). Least Squares Analysis of Data with Unequal Subclass Numbers, Ars 20 - 8, U.S.D.A.

Hemmerle W. J. (1974). Nonorthogonal Analysis of Variance Using - Iterative Improvement and Balanced Residual. Jour. Amer. Statist. Assoc. 69,347.

Hicks R. C. (1973). Fundamental Concepts in the Design of Experiments. John Wiley & Sons, Inc.

Kendall and Stuart (1966). The Advanced Theory of Statistics Vol 3. Third Edition. Griffin.

Kempthorne, O. (1952). The Design and Analysis of Experiments. John Wiley and Sons, Inc., New York.

Krammer C. Y. (1955). On the Analysis of Variance or Two Way - Classification With Unequal Subclass Numbers. Virginia Agricultural Experiment Station. Blacksburg, Virginia.

Lindquist, E. F. (1953). Design and Analysis of Experiments in - Psychology and Education. New York. Hoyhton Mifflin.

Overall y Spiegel (1969). Concernig Least Squares Analysis of - Experimental Data. Psychological Bulletin. Vol 72.

Searle, S. R. (1971). Linear Models, New York. John Wiley & Sons, Inc.

Snedecor, G. W. and Cochran W. G. (1967). Statistical Methods.

Speed, F. M. & Hocking, R. R. (1975). A Full Rank Analysis of Some Linear Model Problems. Jour. Amer. Statist. Assoc. 70,351.

Speed, F. M., Hocking, R. R., Coleman A. T. (1980). Hypothesis to be Tested With Unbalanced Data. Communications in Statistics. A9,2.

Speed, F. M., Hocking, R. R., and Hackney O. P. (1978). Methods - of Analysis of Linear Models With Unbalanced Data. Jour. Amer. - Statist. Assoc. 73, 361.

Yates, F., (1934). The Analysis of Multiple Classifications With Unequal Numbers in the Different Classes. Jour. Amer. Statist. - Assoc.