

CAPITULO 3

DISEÑO y ANALISIS DE EXPERIMENTOS

3.1.PRINCIPIOS DEL DISEÑO DE EXPERIMENTOS

3.1.1.Conceptos generales.

La base del conocimiento de los fenómenos biológicos es el razonamiento inductivo: el observador analiza los hechos observados para lograr una explicación de interés general; una ley general permitirá predecir resultados aún no observados. No obstante, muchas veces las explicaciones logradas son tentativas e hipotéticas, y puede suceder que observadores diferentes proporcionen explicaciones diferentes para un hecho observado. Estos modelos o hipótesis de explicación de la realidad no tienen valor si no concuerdan con esa realidad, una teoría que predice hechos que jamás se observan no es útil. Una hipótesis debe ser comprobada experimentalmente antes de ser aceptada como ley.

Ejemplo 3.1. Los corderos que llegan a la faena provenientes de dos establecimientos pesan diferente. Algunos observadores postulan que las diferencias son genéticas: los corderos más livianos son raza Merino y los del otro establecimiento, que pesan más, son cruza Merino por Texel. Otros técnicos dicen que las diferencias son solo de manejo y los de un establecimiento están mejor alimentados que los del otro. Finalmente, un tercer grupo no descarta que las diferencias observadas sean debidas al azar, y que promediando todos los corderos no existen diferencias. Si los primeros tuvieran razón, corderos de diferente raza tendrán comportamiento diferente con igual manejo, mientras que si los segundos tuvieran razón, esto no sucedería. Se puede realizar un experimento y analizar con cual hipótesis concuerdan los hechos.

Como vemos, el biólogo tiene un auxiliar insustituible en la experimentación: solo observando se puede tomar como cierta una ley explicativa de la realidad. En cierto modo la filosofía es que es tan importante decir lo que es como lo que no es un buen experimento. Aquí analizaremos algunos conceptos de utilidad a los efectos de realizar los ensayos con mejor eficiencia y economía de recursos. Dice Finney (1972, pg. 14) "El estadístico profesional tiene un derecho y una obligación de insistir en la estricta adhesión al significado de los términos técnicos que usa, él está acá en la situación de un zoólogo que objeta el uso de la palabra 'insecto' como incluyendo a las arañas".

Ejemplo 3.2. Un investigador describe la investigación que está llevando a cabo así: "las variables medidas fueron seis (infestación, intensidad, RST, fibra, azúcar por há, peso del tallo) en 3 condiciones diferentes (variedad, lugar y fecha)." Las "variables" que menciona son las variables de respuesta medidas, las "condiciones" (se podrían llamar factores en nuestra terminología) también son variables. Tal vez no haya nada malo en hablar así, ...o tal vez se confunda si alguien le hace referencia a las variables independientes, ya que ese término quedó fuera de su vocabulario. Intentando mejorar la precisión en el hablar comenzaremos el análisis por discutir algunos términos de interés.

Unidad Experimental. Serán los objetos sobre los que se aplica el tratamiento y toman mediciones. Si medimos el crecimiento de corderos frente a raciones diferentes el cordero es la unidad experimental, mientras que si se mide el rendimiento de un potrero por el crecimiento de los corderos que lo pastorean, la parcela es el conjunto de corderos del

potrero. Otra definición de unidad experimental sería la menor subdivisión de material experimental tal que sea posible que diferentes unidades reciban diferentes tratamientos (Cox, 1959). Las unidades experimentales son llamadas parcelas, especialmente en un contexto de producción vegetal.

Unidad de Muestreo. A veces sucede que la unidad muestreada no coincide con la unidad experimental. Dos ejemplos de eso son cuando se mide alguna característica fisiológica de plantas como puede ser largo de pecíolo u otra, donde es imposible medir todas las hojas de todas las plantas de una parcela. Otro ejemplo se presenta en el caso de que la unidad experimental es un conjunto de, digamos, tres animales. Se miden y pesan a los animales individuales, por lo que la unidad de muestreo es el animal.

Error Experimental. Es el error aleatorio que se provoca por la variación del material experimental y la aleatorización que generalmente se aplica para la adjudicación de los tratamientos. El error experimental motiva que parcelas tratadas similarmente proporcionan rendimientos diferentes, dificultando las predicciones.

Error de Muestreo. Es la variación entre unidades de muestreo dentro de una misma unidad experimental. En el caso de los experimentos con animales mencionados, el experimentador puede sentirse tentado de considerar a la variación entre animales como error experimental, siendo que es un error de muestreo.

Factor. Es cada variable experimental independiente. Pueden ser cuantitativos (es decir que se pueden medir) o cualitativos. Los nutrientes fundamentales para las plantas, nitrógeno, fósforo y potasio, son ejemplos de factores cuantitativos de gran interés para nosotros, mientras que razas de animales o variedades de un cultivo, son factores cualitativos.

Nivel de un Factor. Es la intensidad proporcionada a un factor cuantitativo o, en un sentido más amplio, la calidad de un factor cualitativo: 0, 25 y 50 kgs. por há son niveles del factor nitrógeno en fertilización; Calandria, Federal y Colibrí, pueden ser considerados diferentes niveles del factor variedad de trigo.

Tratamiento. Es la combinación de niveles de los factores que se aplica a las parcelas. Si se alimenta con heno de alfalfa a animales estabulados y con praderas de trébol blanco a otros animales, los tratamientos no son la alfalfa y el trébol, como puede parecer, sino "alfalfa en establo" y "trebol blanco a campo", resultando imposible separar los efectos del forraje de los del manejo, los que se dice que están confundidos.

Exactitud. Cercanía entre una observación y la cantidad que se intenta observar.

Precisión. Calidad asociada a una clase de mediciones y referente al modo como concuerdan observaciones repetidas. Para revisar estos dos conceptos ver la sección 1.3.2.

Bloque. Es un conjunto de unidades experimentales parecidas (que tienen algo en común) que se toman en conjunto y en las que se distribuyen los tratamientos. En la terminología que usamos no todo grupo de parcelas es un bloque, sino que tiene que contener todos los tratamientos (bloque completo) y que los tratamientos se hayan aleatorizado separadamente en cada bloque (bloques al azar). Además en el análisis se supone que no interactúan con los tratamientos, de lo contrario no se consideran bloques.

Repetición. Conjunto de parcelas que completen todos los tratamientos. En un diseño en bloques al azar un bloque es una repetición, en un diseño en bloques incompletos pueden ser necesarios varios bloques para hacer una repetición.

Parámetros, Variables, Características. Una visión distingue entre características y las variables que la miden. Por ejemplo una **característica** como precocidad en el ganado se puede medir por **variables** como edad al primer celo, edad al primer parto o tiempo para

llegar a un determinado peso. La precocidad de un cultivo (característica) puede medirse como el tiempo que tarda en florecer la mitad de las plantas (variable). Por otro lado, conviene distinguir entre parámetros y variables. Mediciones como altura y peso de animales, o largo del pecíolo en un cultivo no convienen ser mencionados como parámetros, ya que esta palabra (de origen estadístico) se asocia mejor con características fijas de una población: media, varianzas y correlaciones. Finalmente recordemos que antes distinguimos entre parámetros poblacionales y sus correspondientes estadísticos muestrales, generalmente estimadores de los parámetros (ver capítulo 1).

Los datos biológicos son influídos por un número generalmente grande de factores; ello motiva que aún parcelas cultivadas con un máximo de similitud produzcan de manera diferente. Esta variabilidad ocasiona que al trabajar con datos biológicos, sea difícil determinar lo que se debe al efecto de un factor de interés, y lo que se debe al efecto de los demás factores. Esa variación del azar se ha presentado históricamente con mas claridad en la experimentación biológica que en ninguna otra.

Ejemplo 3.3. Un investigador quiere probar Ralgro, un promotor del crecimiento, en vacas enteras y castradas y dice tener cuatro grupos de vacas: les llama VC, VE, VC' y VE'. VC significa vacas castradas y VE vacas enteras. El considera que las primas (los grupos 3 y 4) son repetición y dice que cada grupo tendrá algunas vacas con Ralgro y otras sin el producto. De acuerdo a nuestra terminología los tratamientos son cuatro:

1- VECR 2- VEsR 3-VCCR 4-VCSR

lo que no quedaba claro en la manera de expresarse del investigador. Por otro lado, él necesita grupos de 4 vacas para probar todos los tratamientos y esos grupos se llaman bloques. Lo que el llama "grupo" es el conjunto de vacas sometidas a un mismo tratamiento y no tiene ningún significado en el léxico del diseño de experimentos. Un punto, tal vez menor, es que cada bloque constituye una repetición (también Fenocchi y Restaino [1988] le llaman repetición a cada bloque -ellos también dicen grupo-adicional).

3.1.2.Requerimientos de un buen diseño.

Se pueden nombrar algunos requerimientos primarios en el diseño de experimentos siguiendo a Cox (1958) como:

i. AUSENCIA DE ERROR SISTEMATICO. Se ha comentado el hecho de que los experimentos comparativos son más usuales y deseables que los experimentos absolutos. Cuando se trata de comparar dos (o más) tratamientos la filosofía es que las parcelas sean iguales en todo menos en el tratamiento. Cualquier otra diferencia entre ellas constituye un error sistemático. Una variante de este requerimiento es que cuando se analizan las diferencias existentes entre un grupo de unidades sometidas a un tratamiento y otro no sometido al mismo debe considerarse la posibilidad de un testigo que sea tratado (en todo menos en el tratamiento en cuestión) como el grupo tratado.

Ejemplo 3.3. Si se desea comparar el efecto de la fertilización potásica en el rendimiento de la caña de azúcar no se puede tomar una zona del campo experimental y fertilizándola concluir que se incremento el rendimiento sin compararla con un testigo del mismo campo y sometido en todo lo posible a las mismas condiciones. Por otro lado, si el experimentador aplica potasio a una parte del campo y deja como testigo a otra, sus conclusiones son puramente descriptivas ya que no demuestra razonablemente la inexistencia de errores sistemáticos.

Ejemplo 3.4. Un ensayo para evaluar el efecto de dos raciones se llevo a cabo en cerdos tomando diez animales para cada tratamiento y comparando los resultados al ser alimentados con ambas en igualdad de condiciones. No obstante, en uno de los tratamientos había 6 machos y en el otro 4. Las conclusiones están viciadas, ya que el efecto de la diferente proporción de machos en las muestras no es descartable que influya en el rendimiento promedio mayor que se obtuvo.

Ejemplo 3.5. Se observa bastante el caso de hacer dos ensayos iguales uno al lado del otro por ejemplo para dos épocas de siembra. El investigador luego quiere comparar las dos épocas de siembra. La comparación tiene valor solamente si el puede asegurar que no existe ninguna diferencia sistemática entre los dos ensayos, cosa que casi seguramente el investigador va a tender a asegurar sin ninguna razón de peso, continuando con su persistencia.

Ejemplo 3.6. Supongamos que a un investigador le quedó:

Tratamiento		1	2	3	4
Año	78	x	x	x	x
		x	x		
Nacimiento	79	x	x	x	x
				x	x

Si hay algún efecto del año de nacimiento, por ejemplo las del año 78 son mejores que las del 79, los tratamientos 1 y 2 se ven beneficiados respecto a los tratamientos 3 y 4.

ii. CUANTIFICACION DE LA INCERTIDUMBRE. El procedimiento de experimentación implica una actitud como "todas las variaciones identificadas deben manejarse de modo que no hagan diferencia entre los tratamientos, las no identificadas deben aleatorizarse y manejarse en términos de probabilidades." Si se logra la inexistencia de errores sistemáticos mencionada antes, los errores remanentes provendrán de factores

aleatorios, si las parcelas se adjudican al azar a los tratamientos. Si los errores son aleatorios se pueden calcular límites de probabilidad entre los cuales se encuentren los verdaderos valores promedio. La cuantificación de la incertidumbre permite decir al investigador "la variedad A tiene 95% de probabilidad de ser mejor que la B".

iii. PRECISION. Cuanto más estrechos sean los límites de confianza mayor es la precisión con la que se está trabajando. El error promedio en las estimaciones de los efectos se conoce como error estándar de esa estimación, y dependerá en su magnitud de: a) La variabilidad de la población de datos. Esa variabilidad se mide a veces con el coeficiente de variación. Es de destacar que el diseño de la experimentación influirá en esta variabilidad. b) El número de observaciones que se tomen dentro de las unidades experimentales y el propio número de unidades experimentales. El error estándar de una estimación es disminuido con el número de repeticiones (unidades experimentales). c) El diseño de experimentos y el método de análisis.

iv. RANGO DE VALIDEZ DE LAS CONCLUSIONES. Todo investigador desea que sus resultados tengan validez en las condiciones de la mayor generalidad posible. En algunos casos eso afecta la esencia del experimento, en otros casos es más discutible la situación. Por ejemplo, en las pruebas de comportamiento de toritos se discute si la superioridad lograda a condiciones de alta alimentación reflejan lo que pasará en condiciones de campo.

v. SIMPLICIDAD. A igualdad de otras características un experimento debe ser lo más simple posible. La razón de este deseo se demuestra por el opuesto: Para qué hacer al experimento innecesariamente complejo?

Ejemplo 3.7. Se ven casos de investigadores que usan 48 tratamientos y se asombran de porque no hay diferencias significativas siendo que el mayor de ellos dio el doble que el menor. Esto es fácil de explicar: algún tratamiento es siempre el mejor y otro es el peor. Cada vez que usamos 48 tratamientos u otro número muy grande de tratamientos la diferencia entre ellos es grande, pero puede no alcanzar a ser significativa. El investigador no debe perder de vista el hecho de que, la no significación implica que, si repitiera el ensayo el que saldrá primero será otro aunque volviera a dar el doble que el de menor rendimiento. Estos son casos en que la complejidad de la situación hace difícil interpretar adecuadamente los resultados. Observemos que no todas las características mencionadas apuntan en el mismo sentido y puede ser necesario un equilibrio entre ellas. Por ejemplo el rango de validez de las conclusiones generalmente se logra a costa de disminuir la simplicidad del experimento.

3.1.3.Principios estadísticos del diseño experimental.

Los métodos de experimentación introducidos a principios de siglo, están basados en tres principios que veremos a continuación. Estos son los principios que figuran en todos los libros de diseños de experimentos.

ALEATORIZACION. Aún cuando las unidades experimentales sean cuidadosamente homogeneizadas, presentan diferencias en su comportamiento. Generalmente, el experimentador no conoce estas pequeñas o grandes diferencias, de modo que adjudicando al azar las parcelas a los tratamientos maneja las diferencias inevitables de un modo que busca igualar la probabilidad de acceso a las mejores parcelas. El sorteo es de importancia fundamental, mediante él se introduce el cálculo de probabilidades en experimentación y se justifica el uso de las pruebas y métodos de análisis que en ese cálculo de probabilidades se basan.

La adjudicación de las parcelas a los tratamientos (véase Hader, 1968 citado por Gill, 1981) puede realizarse de un modo sistemático o subjetivo o al azar. La adjudicación sistemática no nos cubre de algún posible error también sistemático asociado con el criterio y nos imposibilita una estimación del error aleatorio. Subjetivamente se imposibilita la separación entre la influencia del criterio y de los efectos que se analizan.

Ejemplo 3.8. En Inglaterra se realizó una experiencia con la finalidad de estudiar el efecto de la leche pasteurizada como alimento, cuando eso aún era discutible. En cada escuela participante, se alimento a la mitad de los niños con leche pasteurizada y se dejó a la otra mitad como testigo. Student puntualizó que el experimento tenía poco valor ya que las maestras encargadas de la adjudicación de la leche la daban a los niños menos alimentados, con lo que el promedio favorecía a los testigos, que estaban mejor alimentados.

A partir de casos como el anterior se llegó al concepto de la adjudicación aleatoria de los tratamientos a las unidades experimentales, en la máxima extensión posible. Esta aleatorización posibilita:

- a. Estimaciones insesgadas (ausencia de error sistemático) de los parámetros de los modelos explicativos de los sucesos. El error cuadrático medio esta constituido por la precisión más el sesgo, de modo que al decrecer el sesgo se decrece el error.
- b. Cuantificación de la incertidumbre por medio de las pruebas estadísticas que aplican el cálculo de probabilidades.

REPLICACION. Es conocido el concepto de que no se pueden sacar conclusiones fidedignas de pocos datos. En rigor, el conocimiento inductivo solo surge de la interpretación de una regularidad observada de los hechos. Reconocemos dos funciones principales de la replicación que son proveer de una estimación del error experimental y aumentar la precisión.

a) Mayor precisión. Al tomar varias observaciones es de esperar que los factores aleatorios se compensen en alguna medida. Esto se cuantifica cuando al construir intervalos de confianza para la media de una población, tenemos que, si lo realizamos a través de una observación tomada al azar es: $P(Y - 1,96 \sigma_Y < \mu < Y + 1,96 \sigma_Y) = 0,95$ mientras que realizado a través de la media de una muestra aleatoria será:

$$P(\bar{Y} - 1,96 \sigma_{\bar{Y}} < \mu < \bar{Y} + 1,96 \sigma_{\bar{Y}}) = 0,95$$

diferenciándose solamente en la desviación estandar. Como $\sigma = \frac{\sigma}{\sqrt{n}}$ para tamaño de muestras mayores a uno la media muestral proporciona mayor precisión (menor radio al intervalo de confianza) para una confianza dada, en este caso 95%.

b) Posibilidad de estimar el error. La varianza de los datos poblacionales, generalmente llamada del error, no se conoce en el caso general y necesita estimarse. Se puede utilizar una estimación obtenida con otros datos¹, pero evidentemente ello resulta desaconsejable. Es mucho más ventajoso realizar repetición de las observaciones y, con ellas, obtener una estimación de la varianza, lo que a su vez exige un número determinado de repeticiones reafirmando la necesidad de replicar.

Ejemplo 3.9. Debe distinguirse entre la replicación y la medición repetida del mismo resultado. Si varias personas pesan un mismo animal sometido a una ración, es claro que no se obtienen diferentes estimaciones del crecimiento producido por la ración. La varianza de esos datos no mide el error experimental sino el error de medida de los observadores.

Ejemplo 3.10. Considere las dos experiencias siguientes: en una se prueban distintas razas de corderos poniendo animales de ambas razas en tres diferentes potreros; en la otra se prueban densidades de dotación para lo que se pastorea a los animales a distintas densidades en tres potreros. En el primer caso los animales constituyen las unidades experimentales y el tener varios implica replicar. En el segundo caso, cada potrero suministra solamente un dato de ganancia de peso, la variable de interés, de modo que los tres potreros juntos dan una sola replicación del experimento.

CONTROL LOCAL. En casos en que el material experimental se conoce como heterogéneo, ya no resulta admisible que algún tratamiento ocupe por azar las mejores parcelas, sino que debe buscarse la comparación a igualdad de condiciones. Esto conduce a la formación de bloques, grupos de parcelas diferentes entre si pero homogéneos, de tal modo que todos los tratamientos están en cada bloque y su comparación resulte posible con independencia del efecto de cada bloque. Ostle (1965) dice: "En cierto modo control local es sinónimo de diseño experimental". El objetivo fundamental de este principio es mejorar la eficiencia de la experimentación a través de una cuidadosa planificación del muestreo.

Los tres principios estadísticos mencionados trabajan para lograr los objetivos. La aleatorización permite aleatorizar los errores sistemáticos y vuelve posible de cuantificar la incertidumbre haciendo aplicable la teoría de probabilidades. La replicación disminuye la probabilidad de errores sistemáticos por azar, permite estimar el error, aumenta la precisión disminuyendo los límites de confianza y generalmente aumenta el rango de validez porque se aprovecha que se hacen repeticiones para hacerlas en condiciones diferentes de las anteriores (en otros bloques). El control local está destinado primariamente a aumentar la precisión a través del control de los errores sistemáticos y también se aprovecha generalmente para aumentar el rango de validez de las conclusiones haciendo los bloques diferentes entre sí. Notemos que los bloques deben ser homogéneos (las parcelas de un bloque deben ser parecidas entre si) por requerimiento del concepto de bloque, y los bloques se hacen diferentes entre ellos para controlar mejor la variabilidad del material experimental.

¹ En Ingeniería parece que, a veces, se puede extrapolar una variabilidad, pero en biología no.

Como los principios estadísticos buscan facilitar los objetivos.

Aleatorización	Evitar errores sistemáticos
	Cuantificación de la incertidumbre
Replicación	Precisión
Control local	Rango de validez
	Simplicidad

En un estudio el diseño planteado fué el siguiente:

Ejemplo 3.12.”Se usaron 12 vacas para maíz y 12 para sorgo, 3 en cada piquete. Todas las vacas estaban de la 6a semana de lactación en delante y los diferentes grupos tenían igual producción promedio. Se clasificó a las vacas en buenas, intermedias y pobres y se puso una de cada clase en cada grupo”. Notemos la terminología confusa: los grupos son las vacas de un tratamiento, las "clases" a que hacen referencia deberían ser bloques, hubiera sido mejor hacer bloques de 4 vacas y sortear los tratamientos en cada bloque. El hecho de hacer los grupos con igual media, muchas veces implica que cada vez que se incluye una vaca más pesada se incluye una vaca más liviana para compensar. Esa táctica aumenta el error (la varianza dentro de tratamientos).

Un enfoque parece más adecuado: 2 cult x 3 especies x 2 sistemas lo que constituye 12 tratamientos. Para ese factorial se puede usar un factorial común o un diseño en parcelas divididas

S	M	M	S	M	S
CR	RC	CR	CR	RC	RC
Epoca I		Epoca 2		Epoca 3	

3.1.4. Aumento de la precisión.

a. **Métodos técnicos de aumentar la precisión.** (Según Cox, 1958)

a.1. Selección adecuada del material experimental. Siempre que mencionamos este tipo de herramienta viene a la mente casos en los que no se ha cumplido. Por ejemplo el investigador que se plantea un ensayo para comparar los resultados del riego con el cultivo bajo secano, y ese verano resulta extremadamente llovedor. O el otro que proyecta estudiar 5 métodos de control de una plaga que ese año no se presenta. Hay gente que alega que eso no es culpa del investigador, en lo que estamos totalmente de acuerdo, apoyáramos una demanda contra la naturaleza por falta de cooperación, pero no creo que dieran resultado...En cambio hay una cierta posibilidad de influir en los acontecimientos en un caso como el siguiente.

Ejemplo 3.13. Se desea probar cuatro variedades de un cultivo y se eligen los bloques del siguiente modo:

I				
II				
III				
IV				

Vemos que hay parcelas afectadas por engramillamiento, que el investigador no percibe. Cuando hace el análisis no le dan diferencias significativas a pesar de que su impresión es que las variedades difieren en grado importante. Tal vez pueda corregir por una covariable "grado de enmalezamiento" pero la base del problema está en que se poda haber refinado el material experimental.

a.2. Refinamiento de la técnica experimental. Similarmente a lo anterior, se puede a veces aumentar la eficiencia del experimento por un refinamiento de la técnica experimental. Comparando el crecimiento de vacunos es conocido que la variación del contenido del tracto intestinal introduce variabilidad de gran magnitud entre los animales, por lo que se suele evaluarlos luego de un periodo de ayuno.

a.3. Selección de la unidad. Este es el caso que afecta a los investigadores en producción animal que tienen el dilema de decidir cuantos animales ponen en cada unidad experimental o si llevan a cabo un experimento en "*cross over*" o uno en evaluación continua.

a.4. Selección de los tratamientos. Experimentación factorial. Hay un gran potencial de mejorar la eficiencia de un experimento a través de la selección adecuada de los tratamientos. Entre estos los mas mencionados son la experimentación factorial y en estudio de superficies de respuesta el uso de diseños como los centrales compuestos y otros.

La experimentación factorial fue introducida por Fisher y es una de las grandes herramientas de que dispone el investigador actualmente. Dedicaremos un capítulo a los experimentos factoriales mas adelante, pero adelantamos que las principales ventajas que presenta son:

1. El estudio de la interacción que posibilita. Los experimentos a un solo factor deben hacerse a condición fija de los demás, y las conclusiones que se obtienen son solamente válidas para estas condiciones, persistiendo la ignorancia sobre los efectos a otros niveles de factor fijo. Ms importante aún es lo que se observa en casos como cuando el nitrógeno y el fósforo incrementan el rendimiento por separado pero juntos bajan el rendimiento. La única manera de detectar este tipo de fenómeno es por experimentación factorial.

2. Ahorro de recursos (o aumento de la precisión). Probar los efectos del nitrógeno y del riego segun el ejemplo es un gran ahorro de recursos. De no hacerlo así el investigador deberá hacer dos ensayos separados y ello le exigirá 32 parcelas en lugar de las 16 que acá uso, si quiere estudiar cada efecto con la misma precisión que en este caso. El contraste entre las parcelas tratadas con nitrógeno y las no tratadas es analizado como si los diferentes niveles del factor riego constituyeran bloques y viceversa, el contraste entre los niveles de riego es analizado como si los niveles de nitrógeno fueran bloques.

b. Métodos estadísticos de aumentar la precisión.

b.1.Elección del número de repeticiones.

Algunos autores consideran que hay tres tipos de experimentos:

1. “*Screening*” o panorámicos. En ellos el objetivo es "ver que pasa" generalmente al comienzo de una línea de investigación.
2. Experimentos de comprobación de hipótesis. Cuando se tienen elementos para elaborar una hipótesis que interesa demostrar si se cumple o no.
3. Demostrativos. En situaciones en que el investigador ya sabe lo que pasa y lo que intenta hacer con el experimento es demostrarle a los extensionistas y productores el resultado. No interesa el número de observaciones y se hace un par de repeticiones para acallar las críticas.

Es en el segundo tipo de experimento en que interesa calcular un número de repeticiones que cumplir con un determinado objetivo.

b.2.Toma de observaciones adicionales. Uso de covariables.

Es muy común que al comparar variedades interese saber si la diferencia presente se mantiene cuando se descuenta el número de plantas.

Un investigador demostró que los corderos tratados con un producto estimulante del crecimiento de la lana tenían más de los folículos secundarios (que producen lana) que los testigos. Mediante un análisis de covarianza se pudo ver que los corderos ya tenían (por azar) con mas folículos secundarios que los tratados, por lo que no fue efecto del tratamiento.

b.3.Agrupamiento. Uso de diseños experimentales reductores de error.

Dedicaremos el capítulo 4 a discutir los diseños en bloques.

3.1.5.Conclusión.

Muchos investigadores, sobre todo los principiantes y los apurados, expresan el deseo de tener una receta que les aconseje sobre como hacer un buen experimento. Esa receta no existe, al menos nosotros no la conocemos. No obstante no queremos terminar esta sección sin intentar proporcionar un resumen de procedimiento de experimentación. Varios autores (Little y Hills, 1978; Cochran y Cox, 1952; Federer, 1983) han intentado este tipo de resumen, y en general anda alrededor de los siguientes puntos:

- 1.Definición del problema
- 2.Determinación de objetivos
- 3.Análisis de la investigación.
 - 3.1.Evitar vaguedad y excesiva ambición
 - 3.2.Determinar la población objetivo
 - 3.3.Utilización de tratamientos comparativos o absolutos
 - 3.4.Determinar si ser práctico o académico
- 4.Selección de tratamientos
- 5.Selección del material experimental
- 6.Selección del diseño
- 7.Selección de las unidades y del número de unidades
- 8.Control de efectos
- 9.Esbozo del análisis Evitar caer a la Unidad de Biometría con la actitud de Qué me dio?
- 10.Realización
- 11.Análisis en interacción con el estadístico
- 12.Informe

Los puntos 4 a 8 han sido suficientemente discutidos en las secciones previas. Acá corresponde enfatizar los siguientes: la claridad de objetivos es una cualidad generalmente reconocida pero raramente observada. Parece tan obvio y sin embargo se observa ausente o insuficiente tan a menudo. El caso del investigador del ejemplo 3.7. que uso 48 tratamientos es un ejemplo de como agrandar el ensayo conspira contra la claridad de las conclusiones. Los tratamientos comparativos son casi universalmente aceptados como preferibles, pero puede haber excepciones. Niveles de consanguinidad que pueden ser de interés académico no tienen sentido practico ya que animales con esos niveles de consanguinidad no son viables.

Lo mencionado en el punto 3 sobre pensar bien en lo que se va a hacer. Lo siguiente digno de destacar es la importancia de tener un esbozo del análisis antes de la realización del ensayo. Finalmente, consideramos que una investigación no está terminada antes de publicarse los resultados. Para que investigar si no se le dice a nadie o a nadie le interesa? El investigador que no los publica no está cumpliendo con el objetivo de la investigación que es adquirir conocimiento necesario y está induciendo a otros investigadores interesados en el tema a repetir innecesariamente su experimento. Ambas cosas se ven comúnmente en Uruguay.

3.2. ANALISIS DE EXPERIMENTOS EN PARCELAS AL AZAR

3.2.1. Modelo de parcelas al azar.

El describir una variable por su medida de posición, la media aritmética, y su medida de dispersión, la varianza, lleva a pensar en cada valor de la variable como la suma de esa media, y un desvío de la media con esperanza cero y varianza constante: $Y_i = \mu + \varepsilon_i$ de modo que $E[Y_i] = E[\mu + \varepsilon_i] = \mu + E[\varepsilon_i] = \mu$ y $V[Y_i] = V[\mu + \varepsilon_i] = V[\varepsilon_i] = \sigma^2$. Los modelos de diseño experimental consideran a la población de los resultados obtenibles bajo cada tratamiento, como subpoblaciones de una población general de todos los resultados posibles. De este modo, la observación i del tratamiento j : $Y_{ij} = \mu + \tau_j + \varepsilon_{ij}$, se considera la suma de una media general más un efecto del tratamiento j ($\tau_j = \mu_j - \mu$: el efecto del tratamiento j es la diferencia entre la media de la subpoblación constituida por ese tratamiento y la media general, por lo que $\sum \tau_j = 0$) más un desvío de la observación respecto a su media. A este último generalmente lo llamaremos error de predicción, o simplemente error, y engloba todos los factores que no sean contemplados en el modelo y que, por ello, no intervienen en la predicción. Los supuestos del modelo los podemos clasificar en:

- a) Aditividad de los efectos. En caso de no cumplirse el supuesto de que los tratamientos adicionan una cantidad constante τ_j a cada observación, sino que por ejemplo, adicionan un porcentaje, las técnicas de análisis que desarrollaremos no tienen validez an cuando sean las usuales.
- b) Distribución de los errores. Los errores generalmente se suponen aleatorios, acorde con la adjudicación aleatoria de las parcelas; independientes cuando el error que afecta a una parcela no afecta a la vecina; con media cero, es decir que se anulan en promedio; y con varianza común, para todos los tratamientos. Todo ello se puede abreviar en $\varepsilon \sim \text{NID}(0; \sigma^2)$.

3.2.2. Estimación de los parámetros por mínimos cuadrados.

El modelo postula la existencia de parámetros que necesitan estimarse. El método de estimación usado en estos casos es de los mínimos cuadrados, que propone como estimador al valor que haga mínima la suma de cuadrados del error. Ya vimos que el valor esperado de Y_{ij} , la i -sima observación del tratamiento j , es $E[Y_{ij}] = \mu + \tau_j$, las estimaciones de estos parámetros proporcionan una estimación de las medias condicionales para cada tratamiento $\bar{Y}_{ij} = \mu + \tau_j = E[Y_{ij}]$. La diferencia $Y_{ij} - \bar{Y}_{ij} = e_{ij}$ constituye el error de estimación o error experimental, cuya suma de cuadrados queremos minimizar.

$$e_{ij} = Y_{ij} - \bar{Y}_{ij} = Y_{ij} - \mu - \tau_j \quad \sum_i \sum_j e_{ij} = \sum_i \sum_j (Y_{ij} - \bar{Y}_{ij})^2 = \sum_i \sum_j (Y_{ij} - \mu - \tau_j)^2.$$

Por derivación parcial se determinan los estimadores óptimos, como los que anulan la derivada. El estimador de la media general ser $\frac{dSCE}{d\mu} = \sum_k (Y - \mu - \tau) = 0$

$$\text{de donde } \sum \sum Y_{ij} - nk\mu - n\sum T_j = 0 \quad [1]$$

y como suponemos $\sum T_j = 0$ tenemos $\sum_{i=1}^n \sum_{j=1}^k Y_{ij} = nk\mu$ es decir $\mu = \sum_{i=1}^n \sum_{j=1}^k Y_{ij} / nk = \bar{Y}_{..}$.

El estimador de cada efecto de tratamiento se obtiene sucesivamente. Comencemos por el

tratamiento 1. $\frac{dSCE}{dt_1} = \sum \sum (Y - \mu - \tau_1) = 0$, de donde $\sum Y_{i1} - n\mu - n\alpha_1 = 0$

$$\frac{\sum Y_{i1}}{n} = \mu + \alpha_1 \quad [2] \text{ reemplazando } \hat{\mu} = \bar{Y}_{..} \quad \text{y} \quad \hat{\mu}_1 = \bar{Y}_{.1} \text{ y finalmente}$$

$\tau_1 = t_1 = Y_{ij} - \bar{Y}_{ij} = Y_{ij} - \bar{Y}_{.j}$ y, en general, el estimador del efecto del tratamiento j:

$\tau_j = t_j = Y_{ij} - \bar{Y}_{ij} = Y_{ij} - \bar{Y}_{.j}$. Es decir, que cada valor resulta estimado por la media del

tratamiento al que pertenece. Los errores, diferencia entre el valor estimado y el real, son

$e_{ij} = Y_{ij} - \bar{Y}_{ij} = Y_{ij} - \bar{Y}_{.j}$. Las ecuaciones [1] y las del tipo [2], que serán tantas como

tratamientos haya es decir k, constituyen el sistema de ecuaciones normales

$nk \bar{Y}_{..} = \sum \sum Y_{ij} \quad n \bar{Y}_{..} + n t_1 = \sum Y_{i1} \quad n \bar{Y}_{..} + n t_2 = \sum Y_{i2} \quad n \bar{Y}_{..} + n t_k = \sum Y_{ik}$
que, en la práctica, implican que la suma de los errores es nula dentro de cada tratamiento y en el total $\sum \sum e_{ij} = 0, \quad \sum e_{i1} = 0, \quad \sum e_{i2} = 0, \quad \dots \quad \sum e_{ik} = 0$.

Ejemplo 3.14. Ajústense los siguientes datos al modelo $Y_{ij} = \mu + \tau_j + \varepsilon_{ij}$, $\varepsilon \sim \text{NID}(0; \sigma^2)$

Tratamientos	1	2	3	4	Totales
	37	20	41	25	
	31	16	44	22	
	38	15	48	15	
	34	29	47	18	
Totales T_j	140	80	180	80	480
Medias $\bar{Y}_{.j}$	35	20	45	20	30
Efectos t_j	+5	-10	+15	-10	0

Las (estimaciones de) las medias de los tratamientos son: $\mu_1 = \bar{Y}_1 = 35$, $\mu_2 = \bar{Y}_2 = 20$,

$\mu_3 = \bar{Y}_3 = 45$, $\mu_4 = \bar{Y}_4 = 20$. La media de todas esas estimaciones estima la media general

$\mu = \bar{Y}_{..} = 30$, de modo que $t_1 = 35 - 30 = 5$, $t_2 = 20 - 30 = -10$, $t_3 = 45 - 30 = 15$ y $t_4 =$

$20 - 30 = -10$. Como era de esperar, $\sum t_j = 0$. Por último los e_{ij} se calculan como $Y_{ij} - \bar{Y}_{ij} =$

$Y_{ij} - \bar{Y}_{.j}$ quedando la reconstrucción del siguiente modo:

Tratamiento			
1	2	3	4
30+5+2=37	30-10+0=20	30+15-4=41	30-10+5=25
30+5-4=31	30-10-4=16	30+15-1=44	30-10+2=22
30+5+3=38	30-10-5=15	30+15+3=48	30-10-5=15
30+5-1=34	30-10+9=29	30+15+2=47	30-10-2=18
$T_{.1}=140$	$T_{.2}=80$	$T_{.3}=180$	$T_{.4}=80$

$\sum e_{i1}=2-4+3-1=0$, $\sum e_{i2}=0-4-5+9=0$, $\sum e_{i3}=4-1+3+2=0$, $\sum e_{i4}=5+2-5-2=0$

como lo exigen las ecuaciones normales.

3.2.3. Análisis de varianza.

La técnica del análisis de la varianza consiste en separar la variación de un conjunto de datos en partes atribuibles a diferentes causas de variación, de acuerdo a un modelo postulado, y luego compararlas.

PARTICION DE LA SUMA DE CUADRADOS. De acuerdo al modelo, cada observación es la suma de: $Y_{ij} = Y_{..} + t_j + e_{ij}$. De modo que elevando al cuadrado y sumando para los nk valores $\sum Y_{ij}^2 = nkY_{..}^2 + n\sum t_j^2 + \sum e_{ij}^2 + 2$ productos que se anulan y como $\sum Y_{ij}^2 - nkY_{..}^2 = \sum (Y_{ij} - Y_{..})^2$

restando $nkY_{..}^2$ de ambos lados tenemos: $\sum \sum (Y_{ij} - Y_{..})^2 = n\sum t_j^2 - \sum \sum e_{ij}^2$

lo que se puede simbolizar $SC = SCT + SCE$

La suma de cuadrados de tratamientos SCT mide la variación entre las medias de los tratamientos con (k-1) grados de libertad; la suma de cuadrados del error, la variación interna de las muestras con k(n-1) grados de libertad, de modo que los grados de libertad quedaron divididos en correspondencia con las sumas de cuadrados. A los efectos de independizarse del número de observaciones se puede dividir la suma de cuadrados por sus grados de libertad, obteniendo un cuadrado medio (de los desvos)

$$CMT = \frac{SCT}{GLT} = \frac{n \sum t_j^2}{k-1} = \frac{n \sum (\bar{Y}_j - \bar{Y}_{..})^2}{k-1}$$

$$CME = \frac{SCE}{GLE} = \frac{\sum e_{ij}^2}{k(n-1)} = \frac{\sum \sum (Y_{ij} - \hat{Y}_{i,j})^2}{k(n-1)} = \frac{\sum \sum (Y_{ij} - \bar{Y}_j)^2}{nk-k}$$

PRUEBA DE HOMOGENEIDAD DE LOS ESTIMADORES. $H_0: \tau_j = 0$ para todo j (es decir, $\mu_1 = \mu_2 = \dots = \mu_k = \mu$) llamado hipótesis de nulidad, los cuadrados medios constituyen estimaciones de la misma varianza poblacional. En efecto, $\sigma_Y^2 = \frac{\sigma_t^2}{n}$

$$CMT = \frac{SCT}{GLT} = \frac{(\bar{Y}_1 - \bar{Y})^2 + \dots + (\bar{Y}_k - \bar{Y})^2}{k-1} = n\sigma_t^2$$

es un estimador de la varianza por tratamientos, o sea, a través de las medias de los tratamientos. Mientras que:

$$CME = \frac{SCE}{GLE} = \frac{(Y_{i1} - \bar{Y}_1)^2 + \dots + (Y_{ik} - \bar{Y}_k)^2}{n_1 + \dots + n_k - k} = \frac{(n_1 - 1)\sigma_1^2 + \dots + (n_k - 1)\sigma_k^2}{(n_1 - 1) + \dots + (n_k - 1)} = \hat{\sigma}_e^2$$

constituye una estimación de la varianza de la población promediada de k estimaciones, y se la llama estimación de la varianza de la población a través del error. Mientras que la estimación lograda a través de las medias es sesgada, si existe efecto de tratamientos (su esperanza es $\sigma^2 + n(\sum t_j^2 / (k-1))$) la estimación a través de las varianzas internas es insesgada.

Debido a que son independientes, se puede estudiar la homogeneidad de ambas varianzas a través de una prueba F, probando de este modo la hipótesis de nulidad. Estos

resultados se expresan generalmente en un cuadro de análisis de varianza como el que sigue donde expresamos las fórmulas de cálculo y $C = \frac{T^2}{nk}$

Fuente Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	F
Muestra o Trat.	$\sum T_j^2/n - C$	$k - 1$	SCT/GLT	CMT/CME
Error o residuo	por diferencia	$k(n - 1)$	SCE/GLE	
TOTAL	$\sum \sum Y_{ij}^2 - C$	$nk - 1$		

Si el valor de F obtenido es superior al $F(0,95;v_t;v_e)$ de tablas, siendo la probabilidad de tener resultados iguales o mayores a que estos menor al 0,05, rechazamos la hipótesis de nulidad con probabilidad de cometer error tipo I (nivel de significación) menor al 5%.

Ejemplo 3.13. (Continuación). Aplicando la técnica a los datos del ejemplo 3.2 tenemos:

$$C = \frac{(\sum_{i=1}^n \sum_{j=1}^k Y_{ij})^2}{nk} = \frac{T^2}{n}$$

$$SC = \sum \sum Y_{ij}^2 - C = 37^2 + 31^2 + \dots + 17^2 - C = 2.040$$

$$SCT = \frac{\sum_{i=1}^k T_i^2}{n} - C = \frac{35^2 + \dots + 20^2}{4} - C = 1.800$$

$$SCE = SC - SCT = 2.040 - 1.800 = 240$$

de modo que, completando el cuadro, tenemos:

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	F
Muestra o Trat.	1.800	3	600	30**
Error o residuo	240	12	30	
TOTAL	2.040	15		

Se puede comprobar que las fórmulas de cálculo abreviado proporcionan las mismas cantidades que las fórmulas de definición:

$$SC = \sum \sum (Y_{ij} - Y_{..})^2 = (37-30)^2 + (31-30)^2 + \dots + (18-30)^2 = 2.040$$

$$SCT = n \sum (Y_{.j} - Y_{..})^2 = 4[5^2 + (-10)^2 + 15^2 + (-10)^2] = 1.800$$

$$SCE = \sum \sum (Y_{ij} - Y_{.j})^2 = 2^2 + (-4)^2 + \dots + (-2)^2 = 240$$

3.3.CONTRASTE DE MEDIAS

La técnica del análisis de varianza que hemos visto determina significación globalmente. Bastará que una media difiera de las demás para que se encuentre significación. A su vez, si se determina significación, el análisis de varianza no indica si solamente una media difiere de las demás, o son todas diferentes entre sí. Los contrastes individuales permiten un análisis diferente y complementario.

3.3.1.Definiciones.

Contraste de medias es toda función lineal de medias muestrales (suma de productos de medias por coeficientes) tal que la suma de los coeficientes sea nula, o sea

$L = \sum c_j Y_{.j} = c_1 Y_1 + c_2 Y_2 + \dots + c_k Y_k$ es un contraste si $\sum c_j = c_1 + c_2 + \dots + c_k = 0$
Al ser funciones de variables aleatorias (las medias muestrales) los contrastes son variables aleatorias.

Esperanza de un contraste, también llamado a veces verdadero valor del contraste, es $E[L] = E[\sum c_j Y_{.j}] = \sum E[c_j Y_{.j}] = \sum c_j E[Y_{.j}] = \sum c_j \mu_j$

Varianza de un contraste como las medias muestrales son independientes, es

$$V[L] = V[\sum c_j Y_{.j}] = \sum V[c_j Y_{.j}] = \sum c_j^2 V[Y_{.j}] = \sum c_j^2 \frac{\sigma_Y^2}{n}$$

si todas las medias provienen de poblaciones con igual varianza. El estimador de esta varianza, generalmente desconocida, es el cuadrado medio del error que se obtiene por análisis de varianza. Usándolo para obtener una estimación de la varianza del contraste

$$V[L] = V[L] = \frac{\sigma_Y^2}{n} \sum_{j=1}^k c_j^2 \quad \text{y si todas las medias provienen de muestras con igual tamaño,}$$

Distribucion de los contrastes. Como funciones lineales de variables aleatorias normales, los contrastes tienen esa distribución, por lo que pueden estandarizarse:

$\frac{L - E[L]}{\sqrt{V[L]}} = z$ y $\frac{L - E[L]}{\sqrt{V[L]}} = t$ al reemplazar $V[L]$, por su estimación $V[L]$ seguirá la distribución de Student.

Prueba de significacion. Si se desea estudiar la hipótesis $H_0: L = 0$ vs $H_A: L \neq 0$, se puede utilizar este conocimiento de las distribuciones. El criterio de decisión será rechazar

H_0 si, en valor absoluto: $t_0 = \frac{L}{\sqrt{V[L]}} > t$, es decir si $|L| > t \sqrt{V[L]}$. Por esta razón al

producto $t \sqrt{V[L]}$ se le llama Mínima Diferencia Significativa (MDS o LSD por las siglas en inglés).

Intervalos de confianza para contrastes. El conocimiento de la distribución permite también realizar intervalos de confianza para los contrastes de la forma

$$P\{\hat{L} - t \cdot \sqrt{V[\hat{L}]} < L < \hat{L} + t \cdot \sqrt{V[\hat{L}]} \} = P\{\hat{L} - \text{MDS} < L < \hat{L} + \text{MDS}\} = \gamma$$

Cuadrado de un contraste. Cada contraste tiene asociado un cuadrado que se encuentra

por
$$SC(L) = \frac{(nL)^2}{n \sum_{j=1}^k c_j^2}$$
 con un grado de libertad. Cuando los coeficientes son todos 1, la

fórmula se abrevia a
$$SC(L) = \frac{(nL)^2}{nk}$$

Ortogonalidad de contrastes. Dos contrastes $L_1 = \sum c_{1j} Y_{.j}$ y $L_2 = \sum c_{2j} Y_{.j}$ son ortogonales si $\sum c_{1j} c_{2j} = 0$; es decir, si la suma de los productos de los coeficientes correspondientes es nula. La condición de ortogonalidad de contrastes asegura su incorrelación o independencia lineal.

Prueba F para contrastes. Al tener un grado de libertad $SC(L) = CM(L)$ y al dividir por el cuadrado medio del error, se obtienen variables con la distribución $F(1;ve)$, por lo que es idéntica a la variable t en sus resultados con una mecánica opcional para probar la hipótesis de no significación del contraste. La suma de cuadrados de los contrastes constituye la SCT de modo que se habla de descomposición de la SCT por contrastes ortogonales.

En el estudio de los factoriales, y en general en los cálculos que involucran contrastes, se trabaja con totales en lugar de hacerlo con las medias, ya que para muestras de igual tamaño el resultado es el mismo y se facilita el cálculo.

3.3.2.Planteamiento de contrastes.

Con k tratamientos no se puede efectuar más que $(k-1)$ contrastes que sean ortogonales. Pero se pueden construir muchos diferentes grupos de contrastes ortogonales, todos ellos con $(k-1)$ contrastes ortogonales (entendiendo por grupos de contrastes ortogonales a los grupos integrados por contrastes ortogonales dos a dos). Queda a criterio del experimentador elegir los que considere más adecuado a su estudio, con tal de elegirlos previamente a la obtención de los datos, como lo exige la prueba t de Student y Fisher. Tomando los datos del ejemplo 2, plantearemos posibilidades de análisis que pueden darse, de acuerdo a lo que sean los tratamientos, solamente como ejemplos.

3.3.2.5.COMPARACIONES MULTIPLES: LSD y TUKEY. Otro de los casos que surge con mucha frecuencia en el trabajo experimental es la necesidad de contrastar medias que o no se habian pensado comparar antes de la obtención de los datos o forman contrastes no ortogonales. En estos casos no se puede utilizar con rigor la prueba de Student y en su lugar se han propuesto algunos tests alternativos entre los que destacamos los de Tukey y de Duncan.

PRUEBA DE TUKEY. La prueba de Tukey para comparación entre dos medias consiste en determinar una mínima diferencia significativa tal que toda diferencia entre dos medias que sea superior a ese valor se declara significativa. El valor que llamaremos W es: $W = q\sqrt{\frac{1}{2}V[L]}$, donde q es la variable de Tukey que viene tabulada para niveles de significación diferentes con dependencia de los grados de libertad del error y del número de tratamientos que se están analizando. Como los contrastes que se prueban son siempre entre dos medias $\Sigma c_j^2 = 2$ y considerando igual tamaño de muestras (no se aconseja utilizar la prueba con tamaños desiguales de muestras) la fórmula a veces se abrevia en:

$$W = q\sqrt{\frac{1}{2}V[L]} = q\sqrt{\frac{1}{2}\frac{2\sigma_e^2}{n}} = q\sqrt{\frac{\sigma_e^2}{n}} = q\sigma_{\bar{y}}$$

Aplicando la prueba a los datos de nuestro ejemplo 2 tenemos que

$$W_{0,05} = 4,20/.5 = 9,39$$

$$W_{0,01} = 5,50/.5 = 12,30$$

todas medias que difieran en más de esas cantidades son significativamente diferentes por lo que se pueden ordenar de mayor a menor para facilitar el trabajo:

$$\bar{Y}_2 = 20 \quad \bar{Y}_4 = 20 \quad \bar{Y}_1 = 35 \quad \bar{Y}_3 = 45$$

Los resultados se abrevian uniendo con una línea (o poniéndoles una misma letra) a las medias que no son significativamente diferentes, en este caso con una línea llena si son homogéneas al 5% y con una punteada si lo son al 1% solamente.

Existen otras pruebas en las que no nos detendremos pero mencionaremos la de Scheffé, que se utiliza para contrastes complejos, y cuya mínima diferencia significativa se obtiene por $S = \sqrt{(k-1) * F_{(0,95;k-1;gle)} * V[L]}$ donde k es el número de tratamientos del experimento, F es el valor de tablas para el nivel de significación que se desee y $V[L]$ es la varianza del contraste. El lector interesado puede consultar Pimentel Gomes (1978).

Las distintas pruebas proporcionan una opción útil al investigador. La prueba de Tukey es más rigurosa que la de Duncan, es decir que su mínima diferencia significativa es más grande, por lo que una diferencia entre medias (como entre las medias \bar{Y} y \bar{Y} del ejemplo) puede ser declarada significativa por una prueba y no serlo por la otra. En compensación a la menor capacidad de detectar diferencias significativas que tiene Tukey proporciona una menor probabilidad de error tipo I. Las pruebas de Tukey y Scheffé (además de la t de Student, como se recordara) proporcionan utilidad en la confección de intervalos de confianza para los contrastes. La forma de esos intervalos de confianza es: $\bar{L} \pm W$ o $\bar{L} \pm S$

3.3.2.2.UN TESTIGO Y DOS PRODUCTOS, UNO DE ELLOS CON DOS CALIDADES.

Los coeficientes para los contrastes serán:

Efecto	t	a	b ₁	b ₂	L	SC(L)
Test. vs Tratados	-3	+1	+1	+1	-80	$(-80)^2/(4 \times 12) = 133,33$
Producto A vs B	0	-2	-1	-1	-100	$(-100)^2/(4 \times 4) = 416,70$
Diferenc. de Calidad	0	0	-1	+1	100	$(100)^2/(4 \times 2) = 1250,00$

El análisis de varianza será:

Fuente Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	F
(Tratamientos)	(1.800)	(3)	(600)	(30)
Test vs tratado	133,3	1	133,3	4,4ns
Producto A vs B	416,7	1	416,7	21*
Calidad en B	1250	1	1250	62,5*
Error o residuo	240	12	20	
TOTAL	2.040	15		

3.3.2.3.DOS PRODUCTOS Y DOS CALIDADES. Supongamos que tenemos dos productos, por ejemplo suplemento mineral (de dos fuentes) y suplemento proteico (también de dos fuentes diferentes). Entonces nos interesará comparar el comportamiento de los animales con los suplementos minerales y con los suplementos proteicos. Después de eso, nos interesa comparar dentro de los suplementos minerales las dos fuentes estudiadas y también dentro de los suplementos proteicos las dos fuentes. Los coeficientes de los contrastes serán:

Efecto	a ₁	a ₂	b ₁	b ₂	L	SC(L)
A vs B	-1	-1	+1	+1	40	$(40)^2/(4 \times 4) = 100$
Calidad en A	+1	-1	0	0	-80	$(-80)^2/(4 \times 2) = 450$
Calidad en B	0	0	-3	+1	-100	$(-100)^2/(4 \times 2) = 1250$

El análisis de varianza será:

Fuente Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	F
(Tratamientos)	(1.800)	(3)	(600)	(30)
A vs B	100	1	100	5ns
Calidad en A	450	1	450	222,5**
Calidad en B	1250	1	1250	62,5**
Error o residuo	240	12	20	
TOTAL	2.040	15		

Como puede verse a través de los ejemplos, las posibilidades son muchas, y en las secciones siguientes veremos dos de las más importantes. En todos los casos, la suma de los cuadrados asociados a los contrastes totaliza la suma de cuadrados de los tratamientos, verificando que constituyen una descomposición de ésta.

3.3.2.4.EXPERIMENTOS FACTORIALES. Los experimentos factoriales son aquellos en que los tratamientos están constituidos por todas las combinaciones de diferentes niveles de más de un factor. Si los niveles de cada factor son dos, pueden ser dos cantidades diferentes o una de las cantidades ser cero. En general, hablaremos del nivel superior como si implicara presencia y del nivel inferior como sinónimo de ausencia, aunque no siempre sea así. El nivel superior de un factor A será a_1 y el inferior a_0 ; en el ejemplo precedente de nitrógeno y riego n_1 es el nivel superior de nitrógeno (N) y n_0 es el nivel inferior o ausencia de N. Los tratamientos serán, pues, n_0r_0 el testigo que también puede representarse por (1); n_1r_0 el tratamiento con nitrógeno sin riego que puede representarse con n; n_0r_1 o r el tratamiento con riego solamente; y n_1r_1 o nr el tratamiento con ambos. En el cuadro siguiente se presentan los resultados obtenidos, de mismo modo que se presentaron las anteriores descomposiciones de la suma de cuadrados de los tratamientos:

Efecto	(1)	n	r	nr	L	SC(L)
Nitrógeno	-1	+1	-1	+1	40	$(40)^2/(4 \times 4) = 100$
Riego	-1	-1	+1	+1	-80	$(-80)^2/(4 \times 2) = 450$
Interacción	+1	-1	-1	+1	-100	$(-100)^2/(4 \times 2) = 1250$
N sin riego	-1	+1	0	0	40	$(40)^2/(4 \times 4) = 100$
N con riego	0	0	-1	+1	-80	$(-80)^2/(4 \times 2) = 450$
Riego	0	0	-3	+1	-100	$(-100)^2/(4 \times 2) = 1250$

El contraste que indica el efecto del riego se incluye nuevamente en el segundo grupo, solamente para visualizar todo el conjunto de contrastes ortogonales. Los dos grupos de arriba son descomposiciones ortogonales de la suma de cuadrados de los tratamientos que generalmente se eligen, según la interacción esté o no presente. La regla general aconsejada en el análisis de experimentos factoriales es estudiar la interacción, si no hay interacción estudiar los efectos principales y si hay interacción entonces estudiar los efectos simples. El análisis de varianza es:

Fuente Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio F	
(Tratamientos)	(1.800)	(3)	(600)	(30)
Nitrógeno	100	1	100	5ns
Riego	450	1	450	222,5**
Interacción	1250	1	1250	62,5**

Nitrógeno sin riego				
Nitrógeno con riego				
Riego				

Error o residuo	240	12	20	
TOTAL	2.040	15		

La descomposición nitrógeno con y sin riego y nitrógeno e interacción representan diferentes descomposiciones de los mismos dos grados de libertad, lo que explica la permanencia invariable del efecto del riego. Los experimentos factoriales se verán con más detalle en el capítulo 5.

3.3.2.1.NIVELES EQUIDISTANTES DE UN FACTOR. Si los tratamientos representan diferentes cantidades de un factor (por ejemplo, 0, 20, 40 y 60 unidades de nitrógeno por parcela) al investigador le interesa conocer si el incremento es lineal, cuadrático, o cúbico. Los coeficientes para los contrastes que le dan esos efectos se toman de tablas de polinomios ortogonales (por ejemplo, las de Fisher y Yates, 1938)).

Coeficientes para los contrastes:

Efecto	a_0	a_1	a_2	a_3	L	SC(L)
Lineal	-3	-1	+1	+3	- 80	$(-80)^2/(4 \times 20) = 80$
Cuadrático	+1	-1	-1	+1	- 40	$(-40)^2/(4 \times 4) = 100$
Cúbico	-1	+3	-3	+1	-360	$(-360)^2/(4 \times 20) = 1620$

El análisis de varianza será:

Fuente Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	F
(Tratamientos)	(1.800)	(3)	(600)	(30)
Efecto lineal	80	1	80	4ns
Efecto cuadrático	100	1	100	5*
Efecto cúbico	1620	1	1620	81*
Error o residuo	240	12	20	
TOTAL	2.040	15		

3.3.3. Modelos aleatorios.

Ejemplo 3.15. Los siguientes datos son pesos corregidos por edad de progenies de 8 toros (Rovira, 1968).

Toro	183	184	187	190	194	195	198	200
	453	480	447	404	488	460	478	423
	451	470	460	410	454	459	429	405
	475	443	466		534	443	442	449
	402	436	431		511		368	453
	435	347	424				342	467
		435	401					403
		453						
		392						
		380						

Obviamente un modelo adecuado para esos datos debe ser del tipo: $Y_{ij} = \mu + \tau_j + \varepsilon_{ij}$, es decir es un modelo llamado de una sola vía, lo que equivale a un diseño en parcelas al azar. El análisis de varianza para ese modelo y esos datos será:

Fuente Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	F
Toros	21.485,71	7	3.062,24	2,38
Error o residuo	41.130,07	32	1.285,31	
TOTAL	62.565,78	39		

Esta situación tiene algo de especial: acá no interesa saber (probar la hipótesis de) si los toros son diferentes sino estimar la variabilidad que hay entre toros, considerando que los toros son una muestra al azar de una población mayor². Esta situación se conoce como de estimar componentes de varianza y el modelo se conoce como modelo aleatorio.

Componentes de varianza. Una manera de estimar los componentes de varianza es igualando el cuadrado medio con su valor esperado (método de los momentos), otras son por procedimientos como máxima verosimilitud (ML) y máxima verosimilitud residual (REML), ver sección 1.2, página 15. Si usamos el método de los momentos, tenemos $E\{CMT\} = \sigma_e^2$ por tanto, $\hat{\sigma}_e^2 = 1304,92$; $E\{CMT\} = \sigma_e^2 + k\sigma_T^2$ con $k = 4,89$ por tanto: $\hat{\sigma}_T^2 = \frac{3062,24 - 1285,31}{4,8857} = 363,38$.

Coefficiente de correlación intraclase. Consideremos que la varianza del error ($\hat{\sigma}_e^2$) es 1285,31; la varianza entre toros ($\hat{\sigma}_T^2$) es 363,38; la varianza total es 1648,69; entonces la correlación intraclase (sección 2.3.7; pg. 66) es: $r_1 = \frac{363,38}{1648,69} = 0,22$. El autor dice “la variación entre toros resultó significativa, representando el 22% de la variación total”. Notemos que se refiere al coeficiente de correlación intraclase, no al R^2 ($=0,34$) como pudiera pensarse. Otros aspectos de estas situaciones se discutirán más adelante.

² El autor original del análisis hace una prueba de separación de medias (Rovira, pg 17).

3.4. ANÁLISIS DE DATOS DESBALANCEADOS

Para analizar de datos desbalanceados conviene utilizar un procedimiento semejante al utilizado en el capítulo 2, usando la notación matricial. Este procedimiento es más general que lo que hemos venido presentando en este capítulo, pero en caso que los datos sean balanceados los resultados son los mismos. Uno de los pioneros en la utilización práctica en el área agronómica de estos métodos fue Harvey (1971) y nuestra presentación seguirá sus ideas. El procedimiento será: presentar el modelo, formar las ecuaciones normales, resolverlas, hacer el análisis de varianza y los contrastes de interés, así como algún intervalo de confianza. Comenzaremos por discutir los conceptos con el modelo mas sencillo de una sola vía.

3.4.1. Modelo de diseño con un solo factor.

En los modelos de diseño experimental se intenta modelar una variable cuantitativa Y a partir de tratamientos (variables cualitativas). El enfoque de modelos lineales del capítulo anterior sirve para esta situación, pasando las variables independientes a ser indicadoras.

Ejemplo 3.16. Datos artificiales. En la sección anterior cada tratamiento es una subpoblación, lo único que tiene de particular es que la varianza es la misma para todos. La media de cada población puede ser estimada por la media de las parcelas sometidas al tratamiento.

T1	T2	T3
20	31	45
22	29	
24		

Modelo $Y_{ij} = \mu_1 X_1 + \mu_2 X_2 + \mu_3 X_3 + \epsilon$

Donde $X_1 = 1$ si la observación pertenece al tratamiento 1 y
 $= 0$ en caso contrario

$X_2 = 1$ si la observación pertenece al tratamiento 2 y
 $= 0$ en caso contrario

$X_3 = 1$ si la observación pertenece al tratamiento 3 y
 $= 0$ en caso contrario.

Por esta razón le llaman variables indicadoras a las X's. ϵ es un error aleatorio $\epsilon \sim N(0; \sigma^2)$.

Modelo en notación matricial. $\epsilon \sim N(0; I\sigma^2)$.

$$\begin{aligned} 20 &= \mu_1 X_1 + \mu_2 X_2 + \mu_3 X_3 + \epsilon \\ 22 &= \mu_1 X_1 + \mu_2 X_2 + \mu_3 X_3 + \epsilon \\ 24 &= \mu_1 X_1 + \mu_2 X_2 + \mu_3 X_3 + \epsilon \\ 31 &= \mu_1 X_1 + \mu_2 X_2 + \mu_3 X_3 + \epsilon \\ 29 &= \mu_1 X_1 + \mu_2 X_2 + \mu_3 X_3 + \epsilon \\ 45 &= \mu_1 X_1 + \mu_2 X_2 + \mu_3 X_3 + \epsilon \end{aligned}$$

$$Y = \begin{bmatrix} 20 \\ 22 \\ 24 \\ 31 \\ 29 \\ 45 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} + \begin{bmatrix} e11 \\ e12 \\ e13 \\ e21 \\ e22 \\ e31 \end{bmatrix} = X\beta + \epsilon$$

Comentarios. Todo modelo que se pueda escribir en notación matricial $Y = X\beta + \epsilon$ es un modelo lineal. Lo que varían son las matrices.

Si la matriz X tiene solo ceros y unos \rightarrow modelo de diseño experimental.

Si la matriz X puede tomar cualquier valor \rightarrow modelo de regresión.

La matriz puede tomar cualquier valor porque las variables son cuantitativas, por lo tanto la definición de Salvarrey es: Si las variables X son cuantitativas el modelo es de regresión

Si las variables X son cualitativas el modelo es de diseño experimental

¿Y si algunas son cuali y otras son cuanti? El modelo es de covarianza, pero lo vamos a ver en la sección 4.4.

EL MODELO S Y EL PROBLEMA DE LA MEDIA. Hay autores que consideran incluir en el modelo una constante o media general. El modelo resultante se llama modelo S: $Y_{ij} = \mu + \tau_1 X_1 + \tau_2 X_2 + \tau_3 X_3 + \epsilon$

Matricialmente:

$$Y = \begin{bmatrix} 20 \\ 22 \\ 24 \\ 31 \\ 29 \\ 45 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix} + \begin{bmatrix} e11 \\ e12 \\ e13 \\ e21 \\ e22 \\ e31 \end{bmatrix} = X\beta + \epsilon$$

Ahora la suma de las tres últimas columnas es igual a la primera. Por lo tanto la matriz es de rango incompleto. En la práctica eso tiene varias implicancias, computacionales y de interpretación³. Algunas de las estrategias que se usan para superar la situación son, i) si especificamos que las τ son desvíos suman 0, ii) especificar que el último de los tratamientos es cero.

La idea de que los τ sumen cero a su vez puede ser tomada como tal $\sum_j \tau_j = 0$; o que la

suma $\sum_j \mu_j \tau_j = 0$.

¿Quién es μ ? Tomando la primer situación: $\mu = \frac{\sum_i \mu_i}{k}$ en este caso $\mu = \frac{\mu_1 + \mu_2 + \mu_3}{3} =$

$$\frac{22 + 30 + 45}{3} = 32,30$$

Tomando la segunda: $\mu = \frac{\sum_i \mu_i n_i}{\sum_i n_i}$ tenemos $\mu = \frac{3\mu_1 + 2\mu_2 + \mu_3}{3 + 2 + 1} = \frac{171}{6} = 28,50$

Esta última variante nos dice que la estimación de la media general será la media general de los datos, pero tiene el inconveniente que el significado del parámetro depende del número de observaciones, o sea de los datos. En el caso de datos balanceados el resultado es el mismo, por eso nadie se preocupa.

Si el modelo es aleatorio, podemos llegar a situaciones como

$$\hat{\mu}_j = \hat{\mu} + \frac{nh^2}{1 + (n-1)\frac{h^2}{4}} (\bar{Y}_j - \bar{Y}_{..})$$

Por lo tanto μ puede ser llamada media general con todas esas consideraciones presentes.

³ Algunos autores distinguen entre restricciones del modelo y de los parámetros, consideramos que esto es una sutileza.

OTRAS REPARAMETRIZACIONES DEL MODELO. Además de los modelos M y S, existen otras formas de definir los parámetros del modelo, algunos autores les llaman modelos R y Σ .

Modelo Σ - Harvey. Si suponemos que la suma de los efectos de los tratamientos es cero, tenemos que el último es la suma de los otros con el signo cambiado. Entonces en lugar de poner al último ponemos menos uno en todos los demás.

$$\begin{bmatrix} 20 \\ 22 \\ 24 \\ 31 \\ 29 \\ 45 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \end{bmatrix} + \begin{bmatrix} e11 \\ e12 \\ e13 \\ e21 \\ e22 \\ e31 \end{bmatrix}$$

Notemos que estamos tomando como μ a la media de las medias (primera variante anterior en la discusión anterior).

Modelo R, Modelo de rango completo – Mendenhall. Toma como base al primero de los tratamientos y se supone que los otros son diferencias con respecto a ese. El libro de Mendenhall (1968) usa este sistema.

$$\begin{bmatrix} 20 \\ 22 \\ 24 \\ 31 \\ 29 \\ 45 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix} + \begin{bmatrix} e11 \\ e12 \\ e13 \\ e21 \\ e22 \\ e31 \end{bmatrix}$$

δ_1 es la media del tratamiento 1, $\delta_2 = \mu_2 - \mu_1$ la diferencia entre la media del tratamiento 2 y el 1 y $\delta_3 = \mu_3 - \mu_1$ es la diferencia entre la media del tratamiento 3 y el 1.

Versión SAS. SAS dice usar un modelo S, pero al tomar como cero el último de los τ termina siendo equivalente a usar un modelo R así (la μ es la media del último tratamiento):

$$\begin{bmatrix} 20 \\ 22 \\ 24 \\ 31 \\ 29 \\ 45 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix} + \begin{bmatrix} e11 \\ e12 \\ e13 \\ e21 \\ e22 \\ e31 \end{bmatrix}$$

ECUACIONES NORMALES.

forma general **modelo M**

$$\begin{bmatrix} n_1 & 0 & 0 \\ 0 & n_2 & 0 \\ 0 & 0 & n_3 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} T_1 \\ T_2 \\ T_3 \end{bmatrix}$$

ejemplo 1

$$\begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} 66 \\ 60 \\ 45 \end{bmatrix}$$

solución

$$\begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \hat{\mu}_3 \end{bmatrix} = \begin{bmatrix} 22 \\ 30 \\ 45 \end{bmatrix}$$

En el modelo S.

$$\begin{bmatrix} n_{..} & n_1 & n_2 & n_3 \\ n_1 & n_1 & 0 & 0 \\ n_2 & 0 & n_2 & 0 \\ n_3 & 0 & 0 & n_3 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix} = \begin{bmatrix} T_{..} \\ T_{.1} \\ T_{.2} \\ T_{.3} \end{bmatrix}$$

$$\begin{bmatrix} 9 & 3 & 2 & 1 \\ 3 & 3 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix} = \begin{bmatrix} 171 \\ 66 \\ 60 \\ 45 \end{bmatrix}$$

$$\begin{bmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\alpha}_3 \end{bmatrix} = \begin{bmatrix} 45 \\ -23 \\ -15 \\ 0 \end{bmatrix}$$

modelo Σ - HARVEY

$$\begin{bmatrix} n_1 + n_2 + n_3 & n_1 - n_3 & n_2 - n_3 \\ n_1 - n_3 & n_1 + n_3 & n_3 \\ n_2 - n_3 & n_3 & n_2 + n_3 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} T_{..} \\ T_1 - T_3 \\ T_2 - T_3 \end{bmatrix} \quad \begin{bmatrix} 6 & 2 & 1 \\ 2 & 4 & 1 \\ 1 & 1 & 3 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} 171 \\ 21 \\ 15 \end{bmatrix} \quad \begin{bmatrix} \hat{\mu} \\ \hat{\tau}_1 \\ \hat{\tau}_2 \end{bmatrix} = \begin{bmatrix} 32,3 \\ -10,3 \\ -2,3 \end{bmatrix}$$

modelo Mendenhall

$$\begin{bmatrix} n_1 & 0 & 0 \\ 0 & n_2 & 0 \\ 0 & 0 & n_3 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} T_{..} \\ T_1 + T_2 \\ T_1 + T_3 \end{bmatrix}$$

$$\begin{bmatrix} 6 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} 171 \\ 60 \\ 45 \end{bmatrix}$$

$$\begin{bmatrix} \hat{\delta}_1 \\ \hat{\delta}_2 \\ \hat{\delta}_3 \end{bmatrix} = \begin{bmatrix} 22 \\ 8 \\ 23 \end{bmatrix}$$

modelo SAS

$$\begin{bmatrix} n_1 + n_2 + n_3 & n_1 & n_2 \\ n_1 & n_1 & 0 \\ n_2 & 0 & n_2 \end{bmatrix} \begin{bmatrix} \mu_{..} \\ \mu_{.1} \\ \mu_{.2} \end{bmatrix} = \begin{bmatrix} T_{..} \\ T_1 \\ T_2 \end{bmatrix}$$

$$\begin{bmatrix} 6 & 3 & 2 \\ 3 & 3 & 0 \\ 2 & 0 & 2 \end{bmatrix} \begin{bmatrix} \mu_{..} \\ \mu_{.1} \\ \mu_{.2} \end{bmatrix} = \begin{bmatrix} 171 \\ 66 \\ 60 \end{bmatrix}$$

$$\begin{bmatrix} \hat{\mu} \\ \hat{\tau}_1 \\ \hat{\tau}_2 \\ \hat{\tau}_3 \end{bmatrix} = \begin{bmatrix} \bar{Y}_{..} \\ \bar{Y}_{.1} - \bar{Y}_{..} \\ \bar{Y}_{.2} - \bar{Y}_{..} \\ \bar{Y}_{.3} - \bar{Y}_{..} \end{bmatrix}$$

SAS toma como matriz efectiva la mayor de rango completo. Es en esta visión que $\hat{\tau} = t_j$ es el efecto del tratamiento j , la diferencia entre la media del tratamiento y la media general.

ANÁLISIS DE VARIANZA. Dentro de la categoría modelo M, consideremos los siguientes modelos:

$$(1) \quad Y_{ij} = \mu_j + \varepsilon_{ij}$$

$$(2) \quad Y_{ij} = \mu + \varepsilon_{ij}$$

El modelo (1), que postula una media para cada tratamiento, tendrá una SCR con 3 grados de libertad. El modelo (2), que postula una media única para todos los tratamientos, tendrá una SCR con 1 grado de libertad. La diferencia entre ambos (que tendrá 2 grados de libertad) nos dice si vale la pena complicar el modelo considerando tres medias o la media es una sola ($\mu_1=\mu_2=\mu_3$)⁴.

En el ejemplo:

$$\begin{bmatrix} 3 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} T_1 \\ T_2 \\ T_3 \end{bmatrix}$$

donde T_j = total del tratamiento j , espero no confundirlos. Los estimadores son:

$$\begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \hat{\mu}_3 \end{bmatrix} = \begin{bmatrix} \bar{Y}_1 \\ \bar{Y}_2 \\ \bar{Y}_3 \end{bmatrix}$$

La suma de cuadrados de la regresión es, con 3 grados de libertad:

$$SCR = b'X'y = [\hat{\mu}_1 \quad \hat{\mu}_2 \quad \hat{\mu}_3] * \begin{bmatrix} T_1 \\ T_2 \\ T_3 \end{bmatrix} = \sum \bar{Y}_j * T_j = \sum_j^k \frac{T_j^2}{n_j}$$

La SC del modelo con una sola media es: $SCR(0) = b'X'y = [\hat{\mu}][T_{..}] = \frac{T_{..}^2}{nk} = C$, que es el

factor de corrección. La diferencia es: $SC \text{ Tratamientos} = SCT = \sum_j^k \frac{T_j^2}{n_j} - C$ con dos grados

de libertad. Tenemos $(22) (60) + (30) (60) + (45) (45) = 5277,0$. Mientras que $(28,5) (171)=4873,5$ por lo tanto la suma de cuadrados de los tratamientos es:

$$5277,0-4873,5= 403,5$$

En el modelo S la mentalidad es que elegimos entre los modelos (3) y (4) agregando las τ_j :

$$(3) \quad Y_{ij} = \mu + \tau_j + \varepsilon_{ij}$$

$$(4) \quad Y_{ij} = \mu + \varepsilon_{ij}$$

Parece que se vuelven un poco más oscuros los motivos por los que la SCR(3) tiene 3 grados de libertad (son 4 parámetros pero solo 3 son independientes) y la SCR(4) tiene 1. La diferencia sigue sendo 2. El anova es único, sea cual sea la reparametrización.

$$SCR = b'X'y = [\hat{\mu} \quad \hat{\tau}_1 \quad \hat{\tau}_2 \quad \hat{\tau}_3] * \begin{bmatrix} T_{..} \\ T_{.1} \\ T_{.2} \\ T_{.3} \end{bmatrix} = [45 \quad -23 \quad -15 \quad 0] * \begin{bmatrix} 171 \\ 66 \\ 60 \\ 45 \end{bmatrix} = 5277,0$$

⁴ El llamado principio de parsimonia.

Otra curiosidad. Acá presentamos la SCT como: $\frac{\sum_{j=1}^k T_j^2}{n} - C$. También se puede demostrar

que es igual a $\frac{n \sum_{j=1}^k (\bar{Y}_{.j} - \bar{Y}_{..})^2}{n}$.

Contrastes. El análisis de varianza dio significativo, es decir que los tratamientos no son iguales. En el modelo M decimos que las medias no son iguales, en el modelo S equivale a decir que los τ_j no son todos ceros. Correspondería ahora saber que media es diferente de quien. Si nos planteamos dos contrastes como los siguientes:

$$L_1 = \mu_2 - \mu_3$$

$$L_2 = 2\mu_1 - \mu_2 - \mu_3$$

Podemos probarlos tomando la opción **CONTRAST** **t +1 -1** y **t 2 -1 -1**. Esto nos llevará a un prueba F.

Prueba de hipótesis de un solo grado de libertad. Se puede probar por medio de una t de

Student: $t = \frac{\hat{L} - \mu_{\hat{L}}}{\hat{\sigma}_{\hat{L}}}$. Recordemos que una F con un grado de libertad en el numerador

equivale a una t al cuadrado. Por lo tanto ambas pruebas son equivalentes.

Intervalos de confianza. Podemos construir intervalos de confianza para los contrastes y otras funciones usando la distribución de Student, del siguiente modo: $\hat{L}_k \pm t\hat{\sigma}_{\hat{L}}$

LSMeans. Las medias mínimo-cuadráticas son llamadas LSMeans en SAS. Hay gente dice que son la media “que tendría el tratamiento si los datos fuesen balanceados”.

Ortogonalidad de contrastes en caso de datos desbalanceados. La definición que dimos antes de ortogonalidad (sección 3.3.1 página 97) debe ser cambiada para la situación de datos desbalanceados por la siguiente:

$$\sum_{j=1}^k \frac{c_{j1}c_{j2}}{n_j}.$$

	T1	T2	T3	
	522	1243	321	
	3	2	1	
L1	0	+1	-1	
L2	2	-1	-1	
L1	0	+1	-1	
L3	2	-1	-1	

Los contrastes L1 y L2 no son ortogonales según esta definición, los L1 y L3 si lo son.

3.4.2. Segundo ejemplo y los modelos con dos factores.

Ejemplo 3.17. Consideremos los siguientes datos desbalanceados (los datos del ejemplo 1.11 forman parte de estos). Comencemos por analizarlos como 4 tratamientos.

Tratamiento	1	2	3	4
	62	163	60	137
	86	208	62	137
	117	154	72	159
	125	154	75	132
	132	183	52	126
		212		122
		169		136
Totales	522	1243	321	949
Medias	104.4	177.57	64.2	135.57

Modelo. $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$ donde se supone que la j -ésima observación del tratamiento i es la suma de una constante general, un efecto del tratamiento i y un error aleatorio individual de cada observación.

Modelo M.

$$\begin{bmatrix} 62 \\ 86 \\ 117 \\ 125 \\ 132 \\ 163 \\ 208 \\ 154 \\ 154 \\ 183 \\ 212 \\ 169 \\ 60 \\ 62 \\ 72 \\ 75 \\ 52 \\ 137 \\ 137 \\ 159 \\ 132 \\ 126 \\ 122 \\ 136 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix}$$

Modelo S

$$\begin{bmatrix} 62 \\ 86 \\ 117 \\ 125 \\ 132 \\ 163 \\ 208 \\ 154 \\ 154 \\ 183 \\ 212 \\ 169 \\ 60 \\ 62 \\ 72 \\ 75 \\ 52 \\ 137 \\ 137 \\ 159 \\ 132 \\ 126 \\ 122 \\ 136 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{bmatrix}$$

Ecuaciones normales. Las ecuaciones normales son:

Según el modelo M

$$\begin{bmatrix} 5 & 0 & 0 & 0 \\ 0 & 7 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 7 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix} = \begin{bmatrix} 522 \\ 1243 \\ 321 \\ 949 \end{bmatrix}$$

Según SAS

$$\begin{bmatrix} 24 & 5 & 7 & 5 & 7 \\ 5 & 5 & 0 & 0 & 0 \\ 7 & 0 & 7 & 0 & 0 \\ 5 & 0 & 0 & 5 & 0 \\ 7 & 0 & 0 & 0 & 7 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{bmatrix} = \begin{bmatrix} 3035 \\ 522 \\ 1243 \\ 321 \\ 949 \end{bmatrix}$$

Soluciones. Como habíamos notado, con el modelo S las ecuaciones normales son de rango incompleto, por tanto no tienen solución única. Entre las alternativas para superar esta situación, la solución de SAS proviene de anular el último de los coeficientes, en este caso $t_4=0$.

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix} = \begin{bmatrix} 104,40 \\ 177,57 \\ 64,20 \\ 135,57 \end{bmatrix} \quad \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} = \begin{bmatrix} 135,57 \\ -31,17 \\ 42,00 \\ -71,37 \\ 0,00 \end{bmatrix}$$

Si usamos el modelo M no tenemos ese problema, pero tenemos otro.

Análisis de varianza. El análisis de varianza consiste como antes en dividir la variación en partes una atribuible al modelo y la otra al azar: SCR y SCE. La variación total es $SC=y'y$, la SCR = $b'X'y$ y la SCE = $SC - SCR$. Los resultados se presentan en una tabla de análisis de varianza:

Fuentes de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	F
Tratamientos	$b'X'y - C = 40682,53$	3	13560,84	33,04
Error	por diferencia 8209,43	20	410,47	
TOTAL	$y'y - C = 48891,96$	23		

Contrastes. El análisis de varianza se usa primariamente para probar la hipótesis de que los tratamientos son iguales, es decir $H_0: \tau_i=0$ para todo i . Un valor de F significativo indica que: o estamos en presencia de una gran casualidad o los tratamientos no son todos iguales. Si concluimos que los tratamientos no son todos iguales, nos interesa profundizar en esa situación tal como hicimos con datos balanceados en la sección 3.3. Una de las formas de llevar a cabo esa profundización es por medio de contrastes o comparaciones de los tratamientos. Los contrastes en situaciones de datos desbalanceados tienen algunas particularidades.

	T1	T2	T3	T4	
	522	1243	321	949	
Época	+1	+1	-1	-1	
Raza	+1	-1	+1	-1	
Interacción	+1	-1	-1	+1	

	522	1243	321	949	L	
Epoca	1	1	-1	-1	495	
Raza	1	-1	1	-1	-1349	
Interacción	1	-1	-1	1	-93	

	522	1243	321	949	L
Epoca	1	1	-1	-1	
Raza	1	-1	1	-1	
Interacción	1	-1	-1	1	

Intervalos de confianza. Se pueden hacer intervalos de confianza para contrastes y aún para parámetros individuales, recordando que en este último caso la interpretación no es simple.

Pruebas de hipótesis con un solo grado de libertad. Los contrastes son un caso de funciones de los parámetros que tienen un solo grado de libertad, los mas importantes tal vez, pero no los únicos.

Modelo M en el SAS. El SAS no trabaja con modelos M, pero podemos trabajar con los tratamientos y complementamos el análisis por medio de contrastes. Para eso necesariamente tenemos que usar el PROC GLM, ya que es el que tiene la posibilidad de calcular contrastes.

```
DATA ej42; INPUT trat peso; CARDS;
PROC GLM; CLASS trat; MODEL peso=trat;
CONTRAST 'epoca' trat -1 -1 +1 +1;
CONTRAST 'raza' trat -1 +1 -1 +1;
CONTRAST 'epoca*raza' trat +1 -1 -1 +1; RUN;
```

General Linear Models Procedure						
Class Level Information						
		Class	Levels	Values		
		T	4	1	2	3 4
Number of observations in data set = 24						
The X'X Matrix						
INTERCEPT		T 1	T 2	T 3	T 4	PESO
INTERCEPT	24	5	7	5	7	3035
T 1	5	5	0	0	0	522
T 2	7	0	7	0	0	1243
T 3	5	0	0	5	0	321
T 4	7	0	0	0	7	949
PESO	3035	522	1243	321	949	432693
X'X Generalized Inverse (g2)						
INTERCEPT		T 1	T 2	T 3	T 4	PESO
INTERCEPT	0.142857142	-0.142857143	-0.142857143	-0.142857143	0	135.5714285
T 1	-0.142857143	0.3428571429	0.1428571429	0.1428571429	0	-31.1714285
T 2	-0.142857143	0.1428571429	0.2857142857	0.1428571429	0	42
T 3	-0.142857143	0.1428571429	0.1428571429	0.3428571429	0	-71.3714285
T 4	0	0	0	0	0	0
PESO	135.5714285	-31.17142857	42	-71.37142857	0	8209.4285714
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	3	40682.52976190	13560.84325397	33.04	0.0001	
Error	20	8209.42857143	410.47142857			
Corrected Total	23	48891.95833333				
R-Square		C.V.	Root MSE		PESO Mean	
0.832090		16.02116	20.26009449		126.45833333	
Parameter	Estimate	T for H0: Parameter=0	Pr > T	Std Error of Estimate		
INTERCEPT	135.5714286 B	17.70	0.0001	7.65759594		
T 1	-31.1714286 B	-2.63	0.0161	11.86309661		
2	42.0000000 B	3.88	0.0009	10.82947603		
3	-71.3714286 B	-6.02	0.0001	11.86309661		
4	0.0000000 B	.	.	.		
NOTE: The X'X matrix has been found to be singular and a generalized inverse was used to solve the normal equations. Estimates followed by the letter 'B' are biased, and are not unique estimators of the parameters.						
Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F	
EPOCA	1	30468.429762	30468.429762	74.23	0.0001	
RAZA	1	4.725000	4.725000	0.01	0.9156	
EPOCA*RAZA	1	9853.725000	9853.725000	24.01	0.0001	

Interpretación. Como se analizan datos de dos factores cruzados desbalanceados. Primero destaquemos que el procedimiento de SAS para el análisis de datos desbalanceados es el PROC GLM. El GLM ofrece 4 sumas de cuadrados, si los datos son balanceados las 4 son iguales, una manera de medir el desbalance de los datos es, justamente, por la diferencia entre ellas. Las hipótesis docimadas por las 4 sumas de cuadrados de SAS se discuten en varios lugares, uno de los mejores es el manual de SAS. Por otro lado, la suma de cuadrado de la interacción es siempre la misma, lo que cambia es la suma de cuadrados de factores.

Ajustando diferentes modelos. En la práctica del análisis de datos lo primero que el analista hace es verificar si existe interacción o no. Si hay interacción debe recurrir a los efectos simples de los factores, si no la hay, con los efectos principales alcanza.

Una pregunta de ejemplo es: tiene sentido ajustar un modelo con interacción sin efectos principales? Eso equivale a ajustar un modelo del tipo: $Y_{ijk} = \mu + C_{ij} + e_{ijk}$ lo que equivale a un modelo de media de la célula (Hocking, 1984).

DOS FACTORES CRUZADOS SIN INTERACCIÓN.

En el ejemplo que venimos discutiendo los datos son de dos razas de animales que parieron en dos épocas diferentes. Es un caso típico donde se necesita un modelo con dos factores o modelo de dos vías. Vamos a ver como manejar la situación en esos términos.

Modelo. $Y = u + A + B + e$

El modelo puede ser escrito en notación matricial como:

$$\begin{bmatrix}
 1 & 1 & 0 & 1 & 0 \\
 1 & 1 & 0 & 1 & 0 \\
 1 & 1 & 0 & 1 & 0 \\
 1 & 1 & 0 & 1 & 0 \\
 1 & 1 & 0 & 1 & 0 \\
 1 & 1 & 0 & 0 & 1 \\
 1 & 1 & 0 & 0 & 1 \\
 1 & 1 & 0 & 0 & 1 \\
 1 & 1 & 0 & 0 & 1 \\
 1 & 1 & 0 & 0 & 1 \\
 1 & 1 & 0 & 0 & 1 \\
 1 & 0 & 1 & 1 & 0 \\
 1 & 0 & 1 & 1 & 0 \\
 1 & 0 & 1 & 1 & 0 \\
 1 & 0 & 1 & 1 & 0 \\
 1 & 0 & 1 & 1 & 0 \\
 1 & 0 & 1 & 0 & 1 \\
 1 & 0 & 1 & 0 & 1 \\
 1 & 0 & 1 & 0 & 1 \\
 1 & 0 & 1 & 0 & 1 \\
 1 & 0 & 1 & 0 & 1 \\
 1 & 0 & 1 & 0 & 1 \\
 1 & 0 & 1 & 0 & 1 \\
 1 & 0 & 1 & 0 & 1
 \end{bmatrix}
 \begin{bmatrix}
 \mu \\
 \alpha_1 \\
 \alpha_2 \\
 \beta_1 \\
 \beta_2
 \end{bmatrix}
 =
 \begin{bmatrix}
 62 \\
 86 \\
 117 \\
 125 \\
 132 \\
 163 \\
 208 \\
 154 \\
 154 \\
 183 \\
 212 \\
 169 \\
 60 \\
 62 \\
 72 \\
 75 \\
 52 \\
 137 \\
 137 \\
 159 \\
 132 \\
 126 \\
 122 \\
 136
 \end{bmatrix}$$

Ecuaciones normales. Las ecuaciones normales son:

$$\begin{bmatrix}
 24 & 12 & 12 & 10 & 14 \\
 12 & 12 & 0 & 5 & 7 \\
 12 & 0 & 12 & 5 & 7 \\
 10 & 5 & 5 & 10 & 0 \\
 14 & 7 & 7 & 0 & 14
 \end{bmatrix}
 \begin{bmatrix}
 \mu \\
 \alpha_1 \\
 \alpha_2 \\
 \beta_1 \\
 \beta_2
 \end{bmatrix}
 =
 \begin{bmatrix}
 3035 \\
 1765 \\
 1270 \\
 843 \\
 2192
 \end{bmatrix}$$

Este sistema tiene 5 ecuaciones pero rango 3, es decir que necesitamos 2 restricciones para resolver el sistema. Existen diversas propuestas pero SAS maneja la de hacer cero al ultimo de los coeficientes A y al ultimo de los B's.

Restricciones y solución. Las ecuaciones normales reducidas según el procedimiento de SAS son:

	Cuya inversa es:	Los coeficientes son:
$ \begin{bmatrix} 24 & 12 & 10 \\ 12 & 12 & 5 \\ 10 & 5 & 10 \end{bmatrix} $	$ \begin{bmatrix} 24 & 12 & 10 \\ 12 & 12 & 5 \\ 10 & 5 & 10 \end{bmatrix} $	$ \begin{bmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 135,95 \\ 41,25 \\ 0,00 \\ -72,27 \\ 0,00 \end{bmatrix} $

Análisis de varianza. Según el procedimiento de SAS la suma de cuadrados de la regresión es:

$$(135,5714 \ 42 \ -1,3714 \ -1,8) \begin{matrix} 3035 \\ 1765 \\ 843 \\ 522 \end{matrix} = 424483,5$$

Contrastes. Si alguno o ambos factores A o B fueran significativamente diferentes, interesaría saber quien es diferente de quien entre ellos. Eso se puede estudiar por medio de contrastes entre los tratamientos. En el presente ejemplo que es un factorial 2x2 no tiene sentido, ya que son solo dos niveles de cada factor, si el factor es significativo los dos niveles son diferentes.

Intervalos de confianza. También pueden construirse intervalos de confianza para contrastes y otras funciones de las medias de los tratamientos. En la práctica, esta posibilidad se utiliza meno de lo que se podría.

Ajustes secuenciales. Podemos ajustar sucesivamente diferentes modelos. $R(\mu)$, $R(\mu, \alpha)$, $R(\mu, \alpha, \beta)$, $R(\alpha_i | \mu)$, $R(\beta | \mu, \alpha)$

Modelo	Suma Cuadrados	Suma Cuadrad secuencial
$Y_{ijk} = \mu + e_{ijk}$	$R(\mu)$	$R(A \mu)$
$Y_{ijk} = \mu + A_i + e_{ijk}$	$R(\mu, A)$	$R(B \mu, A)$
$Y_{ijk} = \mu + A_i + B_j + e_{ijk}$	$R(\mu, A, B)$	$R(A*B \mu, A, B)$

SAS ajusta 4 sumas de cuadrados. Si los datos son balanceados entonces: $R(A|B)=R(A)$, $R(B|A)=R(B)$ etc. Si los datos son desbalanceados entonces las sumas de cuadrados son diferentes.

3.4.2.3.USO DE SAS. Un programa SAS para el análisis de datos con dos factores sin interacción sería:

PROC GLM; CLASS epoca raza;

MODEL peso=epoca raza/SS1 SS2 SS3 SS4; LSMEANS epoca raza; RUN;

General Linear Models Procedure						
Class Level Information						
Class	Levels	Values				
E	2	1 2				
R	2	1 2				
Number of observations in data set = 28						
The X'X Matrix						
	INTERCEPT	E 1	E 2	R 1	R 2	Y
INTERCEPT	24	12	12	10	14	3035
E 1	12	12	0	5	7	1765
E 2	12	0	12	5	7	1270
R 1	10	5	5	10	0	843
R 2	14	7	7	0	14	2192
Y	3035	1765	1270	843	2192	432693
X'X Generalized Inverse (g2)						
	INTERCEPT	E 1	E 2	R 1	R 2	Y
INTERCEPT	0.113095238	-0.083333333	0	-0.071428571	0	135.94642857
E 1	-0.083333333	0.166666667	0	-1.26883E-17	0	41.25
E 2	0	0	0	0	0	0
R 1	-0.071428571	-1.26883E-17	0	0.171428571	0	-72.27142857
R 2	0	0	0	0	0	0
Y	135.94642857	41.25	0	-72.27142857	0	8214.1535714
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	2	40677.80476190	20338.90238095	52.00	0.0001	
Error	21	8214.15357143	391.15017007			
Corrected Total	23	48891.95833333				
	R-Square	C.V.	Root MSE	Y Mean		
	0.831994	15.63955	19.77751678	126.45833333		
Source	DF	Type I SS	Mean Square	F Value	Pr > F	
E	1	10209.37500000	10209.37500000	26.10	0.0001	
R	1	30468.42976190	30468.42976190	77.89	0.0001	
Source	DF	Type II SS	Mean Square	F Value	Pr > F	
E	1	10209.37500000	10209.37500000	26.10	0.0001	
R	1	30468.42976190	30468.42976190	77.89	0.0001	
Source	DF	Type III SS	Mean Square	F Value	Pr > F	
E	1	10209.37500000	10209.37500000	26.10	0.0001	
R	1	30468.42976190	30468.42976190	77.89	0.0001	
Source	DF	Type IV SS	Mean Square	F Value	Pr > F	
E	1	10209.37500000	10209.37500000	26.10	0.0001	
R	1	30468.42976190	30468.42976190	77.89	0.0001	
T for H0: Pr > T Std Error of						
Parameter	Estimate	Parameter=0	Estimate			
INTERCEPT	135.9464286 B	20.44	0.0001	6.65110680		
E 1	41.2500000 B	5.11	0.0001	8.07413742		
2	0.0000000 B	.	.	.		
R 1	-72.2714286 B	-8.83	0.0001	8.18866991		
2	0.0000000 B	.	.	.		
NOTE: The X'X matrix has been found to be singular and a generalized inverse was used to solve the normal equations. Estimates followed by the letter 'B' are biased, and are not unique estimators of the parameters.						
Least Squares Means						
E	Y					
	LSMEAN					
1	141.060714					
2	99.810714					
R	Y					
	LSMEAN					
1	84.300000					

2 156.571429			
Level of -----Y-----			
E	N	Mean	SD
1	12	147.083333	45.3620954
2	12	105.833333	38.1952718
Level of -----Y-----			
R	N	Mean	SD
1	10	84.300000	29.5711947
2	14	156.571429	28.4921427

MODELO CON DOS FACTORES CRUZADOS CON INTERACCIÓN.

El modelo anterior no contempla la posibilidad de que los factores A y B interactúen. Veamos lo que sucede si intentamos tener en cuenta ese elemento.

Modelo. Un modelo con interacción para los datos presentes es:

$$Y_{ijk} = u + A_i + B_j + C_{ij} + e_{ijk}$$

donde A es el efecto de la poca i (1 o 2), B es el efecto del grupo racial j (también 1 o 2) y C_{ij} es la interacción. El modelo puede ser escrito en notación matricial como:

		Ecuaciones normales. Las ecuaciones normales son:	
62	1 1 0 1 01 0 0 0	$\begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_3 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{21} \\ \gamma_{22} \end{bmatrix}$	$\begin{bmatrix} 24 & 12 & 12 & 14 & 10 & 7 & 5 & 7 & 5 \\ 12 & 12 & 0 & 7 & 5 & 7 & 0 & 0 & 0 \\ 12 & 0 & 12 & 7 & 5 & 0 & 5 & 0 & 0 \\ 14 & 7 & 7 & 14 & 0 & 0 & 0 & 7 & 0 \\ 10 & 5 & 5 & 0 & 10 & 0 & 0 & 0 & 5 \\ 7 & 7 & 0 & 0 & 0 & 7 & 0 & 0 & 0 \\ 5 & 0 & 5 & 0 & 0 & 0 & 5 & 0 & 0 \\ 7 & 0 & 0 & 7 & 0 & 0 & 0 & 7 & 0 \\ 5 & 0 & 0 & 0 & 5 & 0 & 0 & 0 & 5 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_3 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{21} \\ \gamma_{22} \end{bmatrix} = \begin{bmatrix} 3035 \\ 1765 \\ 1270 \\ 843 \\ 2192 \\ 522 \\ 1243 \\ 321 \\ 949 \end{bmatrix}$
86	1 1 0 1 01 0 0 0		
117	1 1 0 1 01 0 0 0		
125	1 1 0 1 01 0 0 0		
132	1 1 0 1 01 0 0 0		
163	1 1 0 0 10 1 0 0		
208	1 1 0 0 10 1 0 0		
154	1 1 0 0 10 1 0 0		
154	1 1 0 0 10 1 0 0		
183	1 1 0 0 10 1 0 0		
212	1 1 0 0 10 1 0 0		
169	1 1 0 0 10 1 0 0		
60	1 0 1 1 00 0 1 0		
62	1 0 1 1 00 0 1 0		
72	1 0 1 1 00 0 1 0		
75	1 0 1 1 00 0 1 0		
52	1 0 1 1 00 0 1 0		
137	1 0 1 0 10 0 0 1		
137	1 0 1 0 10 0 0 1		
159	1 0 1 0 10 0 0 1		
132	1 0 1 0 10 0 0 1		
126	1 0 1 0 10 0 0 1		
122	1 0 1 0 10 0 0 1		
136	1 0 1 0 1 0 0 1		

Restricciones y solución. Las ecuaciones normales reducidas segun el procedimiento de SAS son:

	Cuya inversa es:	Los coeficientes son:
$\begin{bmatrix} 21 & 12 & 14 & 7 \\ 12 & 12 & 0 & 0 \\ 14 & 0 & 14 & 0 \\ 7 & 0 & 0 & 7 \end{bmatrix}$	$\begin{bmatrix} 0,1429 & -0,1429 & -0,1429 & 0,1429 \\ -0,1429 & 0,2857 & 0,1429 & -0,2857 \\ 0,1429 & 0,1429 & 0,3428 & -0,3428 \\ 0,1429 & -0,2857 & -0,3428 & 0,6857 \end{bmatrix}$	$\begin{bmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\beta}_1 \\ \hat{\gamma}_{11} \end{bmatrix} = \begin{bmatrix} 135,57 \\ 42,00 \\ -71,37 \\ -1,80 \end{bmatrix}$

		Los coeficientes son:
		$\begin{bmatrix} 135,57 \\ 42,00 \\ 0,00 \\ -71,37 \\ 0,00 \\ -1,80 \\ 0,00 \\ 0,00 \\ 0,00 \end{bmatrix}$

Análisis de varianza. Según el procedimiento de SAS la suma de cuadrados de la regresión es:
$$\begin{bmatrix} \hat{\mu} & \hat{\alpha}_1 & \hat{\beta}_1 & \hat{\gamma}_{11} \end{bmatrix} \begin{bmatrix} T_{..} \\ T_{1.} \\ T_{.1} \\ T_{11} \end{bmatrix} = \begin{bmatrix} 135,57 & 42,00 & -71,37 & -1,80 \end{bmatrix} \begin{bmatrix} 3035 \\ 1765 \\ 843 \\ 522 \end{bmatrix} = 424483,5$$

Análisis de la varianza según el modelo S. De acuerdo con la filosofía del modelo S, el análisis de la varianza consiste en estudiar sucesivamente modelos cada vez más sencillos:

Modelo	Suma Cuadrados	Suma Cuadrad secuencial
$Y_{ijk} = \mu + e_{ijk}$	$R(\mu)$	
$Y_{ijk} = \mu + A_i + e_{ijk}$	$R(\mu, A)$	$R(A \mu)$
$Y_{ijk} = \mu + A_i + B_j + e_{ijk}$	$R(\mu, A, B)$	$R(B \mu, A)$
$Y_{ijk} = \mu + A_i + B_j + A*B_{ij} + e_{ijk}$	$R(\mu, A, B, A*B)$	$R(A*B \mu, A, B)$

Un comentario es que no tiene sentido ajustar un modelo con interacción y sin efectos principales (los parámetros d

3.4.3. Un tercer ejemplo. La discusión anterior fue más computacional. Podemos intentar un enfoque más interpretativo de la situación a través de estos ejemplos. Supongamos que tenemos dos factores A y B, el primero con dos niveles y el segundo con tres niveles.

Ejemplo 3. Observemos la siguiente tabla, donde tenemos dos lugares (factor A) y tres variedades (factor B).

	B1	B2	B3	
A1	μ_{11}	μ_{12}	μ_{13}	$\mu_{1.}$
A2	μ_{21}	μ_{22}	μ_{23}	$\mu_{2.}$
	$\mu_{.1}$	$\mu_{.2}$	$\mu_{.3}$	$\mu_{..}$

EL MODELO M. El efecto principal del factor A tiene un grado de libertad, el efecto principal del factor B tiene 2 grados de libertad y la interacción tiene dos grados de libertad.

Efecto de A. $A = \mu_{1.} - \mu_{2.}$ El efecto principal del factor A (localidad) es la diferencia que existe entre el rendimiento promedio en ambas localidades.

Efecto de B. El factor B tiene tres niveles (tres variedades) por lo tanto el efecto de B tiene dos grados de libertad. Esto quiere decir que un conjunto de contrastes para medir el efecto

del factor B tiene dos integrantes: $B = \begin{bmatrix} \mu_{.1} - \mu_{.2} = 0 \\ \mu_{.1} - \mu_{.3} = 0 \end{bmatrix}$.

Interacción. La interacción puede ser medida por el siguiente conjunto de contrastes:

$A*B = \begin{bmatrix} \mu_{11} - \mu_{12} - \mu_{21} + \mu_{22} = 0 \\ \mu_{11} - \mu_{13} - \mu_{31} + \mu_{23} = 0 \end{bmatrix}$. También la interacción tiene dos grados de libertad.

Ejemplo 4. Supongamos que falta una celda, como en el siguiente caso en que suponemos que la celda que falta es la 23. Entonces el segundo contraste de la interacción no se puede calcular, por lo tanto la interacción tiene un solo grado de libertad. El SAS se confunde con eso y dice que el contraste no es estimable.

	B1	B2	B3	
A1	10	12	17	13
A2	11	13		12
	$\mu_{.1}$	$\mu_{.2}$	$\mu_{.3}$	$\mu_{..}$

Pero no solo eso, también los efectos principales cambian su sentido. Tenemos dos estrategias: 1) considerar como efecto del factor A a la media de la primera fila contra la media de la segunda o 2) considerar las cuatro celdas completas. ¿Cuál deberemos usar?

Si seguimos la estrategia 1 vemos que la media de la localidad 1 es 13 y la media de la localidad 2 es 12, por lo tanto el investigador pensará que la localidad 1 es mejor. Pero si miramos el cuerpo de la tabla vemos que ¡las dos variedades que estuvieron en ambas localidades rinden mas en la localidad A2!

Mi intención no es promover la estrategia 2, sino alertar sobre la situación y estimular al investigador que medite sobre sus datos.

LA INTERPRETACIÓN SIGUIENDO EL MODELO S

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

Si trabajamos con el modelo S nos postulamos un modelo y nos construimos las ecuaciones normales, pero estas no tienen solución única.

Resolución de las Ecuaciones Normales. Necesitamos una restricción para A, otra para B, y una para C a cada nivel de A e B.

Estimación Puntual. Funciones Estimables.

La estimación puntual en el modelo S lleva a los problemas de no estimabilidad de algunos parámetros. Son estimables: los valores esperados del modelo y las funciones de ellos. Por ejemplo: son estimables $\mu + A_i$; $\mu + A_i + B_j$; $\mu + A_i + B_j + A*B_{ij}$ y no son estimables; A_i ; B_j y $A*B_{ij}$. Entonces tenemos equivalencia con el modelo M: las medias de celdas son estimables.

Test de Hipótesis. En caso de duda se usa la suma de cuadrados tipo III.

Más detalle sobre el análisis de varianza. Ajustes secuenciales: $R(\mu)$, $R(\mu, \alpha)$, $R(\mu, \alpha, \beta)$, $R(\mu, \alpha, \beta, \gamma)$, $R(\alpha_i | \mu)$, $R(\beta_j | \mu, \alpha)$, $R(\gamma_{ij} | \mu, \alpha, \beta)$

Fontes Variação	Soma Quadrados	Graus Liberdade	Quadrado Médio	F
μ				
1				
$\alpha \mu$				
1				
(μ, α)				
2				
B1 vs B2				
1				
B1 B2 vs B3 1				
$(B \mu, \alpha)$				
2				
μ, α, B				
4				
Residuo				
2				
TOTAL				

3.4.4. Casos Mas Complejos.

MODELOS CON TRES O MAS FACTORES. Hasta ahora hemos estado discutiendo modelos con dos factores o modelos de dos vías. En muchos casos los modelos deben contemplar la existencia de mas de dos factores, por ejemplo en un cuadrado latino. El modelo para cuadrado latino, que será discutido en el próximo capítulo, contempla tres factores: $Y_{ijk} = \mu + A_i + B_j + T_k + e_{ijk}$. Un cuadrado latino es, por su naturaleza, balanceado pero sobre la base de ese modelo de tres factores o tres vías se pueden presentar otras situaciones.

Ejemplo 3.16. Un modelo con tres factores se presenta en mucho en la investigación con datos lecheros. Son frecuentes modelos de la forma: $Y = \mu + H + Es + Ed + Año$, donde Y es la producción de leche en una lactancia de 305 días de una vaca, H es rodeo en el que se produjo la lactancia, Est es la estación de parición, Ed es la edad de la vaca al parto y Año es el año (cronológico) en que la vaca parió.

Ejemplos 3.17. Otro modelo usado con vacas lecheras es: $Y = \mu + A + M + A*M + e$, donde A es el año de la parición, M es el mes de parición, y A*M es la interacción entre el año y el mes de parición.

$Y = \mu + A + Es + A*Es + e$, donde A es el año, Es es la estación de parición, y de nuevo A*Es es la interacción entre año y estación de parición.

$Y = \mu + A + Es + A*Es + EM + Es*EM + e$, donde al modelo anterior le agregamos EM que es un efecto de la edad de la madre, y su interacción con la estación de parición. Este modelo se ha usado bastante para factores de corrección en peso al destete de terneros en razas de carne.

$Y = \mu + HYS + Toro + e$, es un modelo que se ha usado mucho en la evaluación genética de toros.

Normalmente en este tipo de estudio se incluyen interacciones entre algunos de los factores en estudio y se especifica como fijos o aleatorios a los diferentes factores. Este último punto no será discutido mucho en este libro, pero algunos elementos se verán en la sección 5.4.

DOS FACTORES ANIDADOS. Si los factores están anidados y no cruzados, la situación tiene sus diferencias. Siempre suponiendo un modelo en que todos los factores son fijos.

Ejemplo 3.18. Datos de producción de leche a 305 días. En el siguiente ejemplo tenemos hijas de un toro (T1) con tres vacas (162, 186 y 317) y de otro toro (T2) con dos vacas diferentes (225 y 132). Medimos, por ejemplo, la producción de leche de hijas de esos cruzamientos en la primera lactancia.

Toro	T1			T2	
Vaca	162	186	317	225	132
	3659	3168	2934	6695	6335
	3276	2940	2265	6570	6145
Totales	6935	6108	5199	13265	12480
Medias	3467,5	3054,0	2599,5	6632,5	6240,0

Un modelo que representa la situación es el siguiente: $Y_{ijk} = \mu + T_i + V_{ij} + e_{ijk}$. Podríamos numerar a los efectos de vacas como 1, 2, ... dentro de los efectos de cada uno

de los toros. Las vacas 1 y 2 del primero toro (162 y 186) no corresponden con las vacas 1 y 2 del segundo toro (que son las 325 y 132) como sera si los efectos estuvieran cruzados. Se dice que el efecto de la vaca está anidado dentro del efecto toro o también se les dice jerárquicos. Los valores serán identificados con dos subíndices: el primero indica el padre (1 o 2 acá) y el segundo el orden dentro de cada toro. Así la vaca 162 es la V11, la vaca 225 es la V21, etc.

Un toro T1 inseminó a dos vacas V1 y V2, otro toro T2 inseminó a tres vacas V3, V4 y V5. **Dos factores anidados vistos según el modelo M.** En el modelo M las celdas son la unidad, y las medias de celda los estimadores, o en otras palabras cada celda es una subpoblación. Por lo tanto, el análisis aconsejado es manejar los efectos como contrastes. Veamos el ejemplo anterior. Tenemos 5 celdas (tratamientos), si nos interesa comparar los toros (“Efecto toro”) hacemos el contraste L1 entre los tratamientos (1 y 2) con (3, 4 y 5). Para comparar las vacas dentro de toro 1, usamos el contraste L2. Los efectos de vacas dentro del toro 2 se miden por dos contrastes (dos grados de libertad). En resumen:

	Toro 1		Toro 2		
	Vaca 1	Vaca 2	Vaca 3	Vaca 4	Vaca 5
L1=Efecto Toro	+3	+3	-2	-2	-2
L2= Efecto B	+1	-1	0	0	0
L3=B(T1)	0	0	+1	-1	0
L4=B(T2)	0	0	+1	0	-1

Parece ridículo hablar de “vacas anidadas en toros”, pero creo que nos entendemos.

Dos factores anidados vistos según el modelo S. Si hacemos el análisis con un modelo S, como usa el SAS. Estamos pensando en un modelo como el siguiente:

$Y_{ijk} = \mu + \alpha_i + \beta_{ij} + e_{ijk}$ algunos autores escriben $\beta_{j(i)}$ pero no parece esencial.

El análisis de varianza sería el siguiente.

Fuente de Variación	Suma Cuadrados	Grados Libertad	Cuadrado Medio	F
Toros	$R(\alpha_i \mu)$	1		
Vacas(Toros)	$R(\beta \mu, \alpha)$	3		
Error				
TOTAL	SC	4		

Reducciones secuenciales $R(\mu)$, $R(\mu, \alpha_i)$, $R(\mu, \alpha_i, \beta_j)$ por lo tanto: $R(\alpha_i | \mu) = R(\mu, \alpha_i) - R(\mu)$, $R(\beta | \mu, \alpha) = R(\mu, \alpha_i, \beta_j) - R(\mu, \alpha_i)$ $SC = R(\mu, \alpha_i, \beta_j) - R(\mu)$.

5. Test de Hipótesis.

Las hipótesis que interesa docimar son si los toros presentan diferencias entre si y si las vacas de un toro difieren (H : Todos os B_i son iguales es equivalente con $R(\beta | \mu, \alpha)$).

CASOS MAS COMPLEJOS DE FACTORES CRUZADOS Y ANIDADOS.

Algunos factores pueden estar cruzados y otros anidados, como vemos en el siguiente ejemplo.

Ejemplo 3.19. Un ejemplo complejo con factores cruzados y anidados del área de producción lechera muy frecuente es de la siguiente forma. Si tenemos 5 toros, dos canadienses y tres estadounidenses, que han tenido hijos en dos rodeos, el modelo es $Y_{ij} = u + H_i + G_j + T_{k(j)} + e_{ijk}$, donde H_i es el efecto del rodeo, G_j es el efecto del grupo j de toros (de origen estadounidense o canadiense) y $T_{k(j)}$ es el efecto del toro k dentro del grupo j . Notemos que $i=1,2$ $j=1,2$ y $k_{j(1)}=1,2$ mientras que $k_{j(2)}=1,2,3,4,5$.

MODELOS DE COVARIANZA. Si algunas variables son cuantitativas, el modelo es considerado de covarianza. Muchas veces se consideraba en estadística experimental que la covariable (la variable cuantitativa del modelo) era una molestia y el llamado análisis de covarianza se orientaba a corregir por el efecto de la covariable (Steel & Torrie, 1980; Snedecor, 19xx). Algunas veces la variable cuantitativa es tan o más importante que la cualitativa (Dobson, 1993, pag. Xx).

Ejemplo 3.21. Rovira y factores de corrección

$$Y = u + \text{Sexo} + \text{Edad} + \text{Toro} + B X + e$$

MODELOS DE COMPONENTES DE VARIANZA. Supongamos que tenemos un modelo con dos factores. Si los factores son fijos, el modelo se llama fijo, si los factores son aleatorios el modelo también lo es, y si uno de los factores es fijo y el otro aleatorio el modelo se denomina mixto. Si los términos del modelo son fijos se llaman parámetros y lo que el investigador busca es estimarlos. Si son aleatorios el objetivo lógico es estimar la varianza del factor. Los modelos para este tipo de situación se conocen como modelos de componentes de varianza. Estimar componentes de varianza con datos desbalanceados es complejo y ha dado lugar a muchos estudios en la literatura estadística. En modelos de una sola vía la situación no se complica tanto, como se vio en la sección 3.3.3, pero en modelos con mas factores comienza la complejidad.