Diseños Factoriales

Daniel Sibaja Salazar

Contents

Detalles 2

Interacción **2**

Un experimento sin interacción. 2

Un experimento sin interacción. 3

Análisis de varianza. **4**

Con interacción. 4

Sin interacción. 4

Forma de calculo. 4

De interacción. 4

Variación de los errores 5

Análisis de Varianza. **5**

Promedios estimados. (Ignorando la interacción) 5

Promedios Observados 6

ANOVA (Ignorando interacción) 6

Parametrizaciones del modelo **6**

Suma Nula 6

Tratamiento de referencia 7

Comparaciones múltiples **8**

Ejemplo del modelo sin interacción 9

Variables confusoras **9**

A partir de ahora vamos a ver experimentos con más de un factor. Por lo cual hay que tomar en cuenta nuevos conceptos.

En este sentido a cada observación se le aplican tratamientos que son el resultado de la combinación más de un factor. (En el curso generalmente dos)

Nos interesa medir los efectos sobre la variable respuesta de dos o más factores y la interacción entre ellos.

Detalles

Si tenemos que un factor denominado A y otro denominado B, la cantidad de tratamientos es AxB, la combinación de todos los niveles de A con todos los de B.

Si el experimento es balanceado tenemos que n=TratamientosxReplicas, es decir n=AxBxr

Interacción

La interacción podemos definirla como cuando el efecto de una variable independiente sobre la variable respuesta depende del nivel o de la presencia de otra variable independiente.

NOTA: es de suma importancia verificar si hay interacción *antes* de hacer cualquier análisis

Un experimento sin interacción.

Cuando no hay interacción los efectos de ambos factores *pueden analizarse de forma separada*, analizando el promedio del nivel del factor $\bar{y}_{i.}$ o $\bar{y}_{.j}$ respecto al promedio general $\bar{y}$

Recordemos el caso de un factor:

$$\tau_i = \mu_j - \mu$$

Entonces en el caso de dos factores es igual, pero los niveles tienen "otros nombres"

$\bar{y}_{ij}$ = es el promedio del nivel i del factor A en el nivel j del factor B

También es posible determinar:

- $\bar{y}_{i.}$ es el promedio de del factor A sin verlo para un nivel específico de B.
- $\bar{y}_{.j}$ es el promedio de del factor A sin verlo para un nivel específico de B.

Efectos con 2 factores (con interacción)

- **Factor A:** α_i es el i-ésimo nivel del factor A.
  - Es decir: $\alpha_i = \mu_{i.} - \mu$
  - O bien estimado: $\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}$
- **Factor B:** β_j es el j-ésimo nivel del factor B.
  - Es decir: $\beta_i = \mu_{.J} - \mu$
  - O bien estimado: $\hat{\beta}_i = \bar{y}_{.j} - \bar{y}$

Nota: La suma de los efectos deben ser iguales a 0.

Nota: Si no es balanceado hay que hacer promedios ponderados

En resumen, cuando no hay interacción:

El efecto de un nivel del factor A es el mismo para cada nivel del factor B.

El modelo sin interacción

$$\mu_{ij}^{SI} = \mu + \alpha_i + \beta_j$$

También se pueden conseguir los efectos simples de cada nivel

$$\alpha_i = \mu_i. - \mu$$

$$\beta_j = \mu_j. - \mu$$

Un experimento sin interacción.

Los efectos de un factor dependen del nivel en el que nos encontremos del otro factor.

En este caso ya *NO* se puede hablar del efecto de un nivel del factor A porque depende de en cual nivel del factor B estoy.

NOTA: Cuando se prueba que no hay interacción se debe de ajustar un nuevo modelo, con diferentes grados de libertad, quiere decir que los cuadrados medios puede cambiar (el residual)

El modelo con interacción.

$$\mu_{ij}^{CI} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

Tomando esto en cuenta la diferencia entre el modelo con interacción y sin interacción es la siguiente

$$(\alpha\beta)_{ij} = mu_{ij}^{CI} - \mu_{ij}^{SI}$$

Este modelo produce estimaciones de las medias por tratamiento iguales a las medias observadas de cada tratamiento, debido a que siempre se está bajo el efecto de la interacción (aunque este sea 0)

$$\mu_{ij}^{CI} = \bar{y}_{ij}$$

En este caso es necesario hacer una prueba de hipótesis para saber si hay interacción o no.

$$H0 : (\alpha\beta)_{ij} = 0$$

Estimación de efectos de interacción Estas estimaciones pueden ser obtenidas mediante la estimación de las medias de ambos modelos

Recordemos

$$\mu_{ij}^{SI} = \mu + \alpha_i + \beta_j$$

Entonces podemos estimar los efectos de la interacción así

$$(\alpha\beta)_{ij} = mu_{ij}^{CI} - \mu_{ij}^{SI} = \bar{y}_{ij} - (\bar{y} + \alpha_i + \beta_j)$$

Los efectos de interacción deben todos sumar 0, si se colocan en una tabla, deben de sumar 0 entre filas y entre columnas.

En muchos casos para tener acceso a la tabla completa, basta con tener (gl) efectos. También los grados de libertad son los coeficientes de interacción que tendremos (en caso de necesitarlos)

Los gl se consiguen de esta forma.

$$(a-1)*(b-1)$$

Hay que tomar en cuenta también que en el caso con interacción existen efectos de interacción y efectos del factor. Si hay interacción **dependen** del nivel del factor en el que se esté.

Análisis de varianza.

Con interacción.

La lógica e la misma que con un factor, solo que ahora con interacción se divide la variabilidad total (Suma de cuadrados Total) en cuatro distintas fuentes de variación.

- Variación *entre* los promedios del primer factor.
- Variación *entre* los promedios del segundo factor.
- Variación *debida* a la interacción.
- Variación *dentro* de los tratamientos.

Sin interacción.

La suma de cuadrados total se descompone de la siguiente forma.

- Variación *entre* los promedios del primer factor.
- Variación *entre* los promedios del segundo factor.
- Variación *dentro* de los tratamientos.

NOTA: La suma de cuadrados total no cambia. Solo se desglosa de diferente forma, tampoco cambia la variación entre promedios del primer factor ni del segundo. Solo cambia la variación dentro de los tratamientos, ya que es la que se desglosa con la variación debida a la interacción.

Forma de calculo.

Entre los promedios de los factores.

$$CMFactor1 = \frac{\sum r_{i.}(\bar{y}_{i.} - \bar{y})^2}{a-1}$$

$$CMFactor1 = \frac{\sum r_{.j}(\bar{y}_{.j} - \bar{y})^2}{b-1}$$

De interacción.

$$CMInt = \frac{\sum r(\hat{\alpha\beta})_{ij}^2}{(a-1)(b-1)}$$

Variación de los errores

Es la diferencia entre cada valor observado y la media del tratamiento a la que pertenece, el residual es lo mismo pero respecto a la media estimada de cada tratamiento.

Cuanto se desvía cada valor del promedio del tratamiento.

El cuadrado medio residual es una medida de la variabilidad de los residuales en conjunto.

Con o sin interacción debería cambiar, no necesariamente debe de ser mayor o menor.

Cuando no se usa interacción (porque no se rechaza la hipótesis) los grados de libertad aumentan y esto puede llevarnos a una mejor estimación.

$$CMRes = \frac{\sum_i \sum_j (r_{ij} - 1) s_{ij}^2}{n - k}$$

donde:

- n= tamaño de muestra total
- k= número de tratamientos
- r_{ij}= Réplicas por tratamiento

Este valor es deseable que sea pequeñito.

Para el caso sin interacción la SCRes cambia un poco.

$$CMRes^{SI} = \frac{SCInt + SCRes^{CI}}{(a-1)(b-1) + (n - ab)}$$

Análisis de Varianza.

Ahora tenemos tres hipótesis en lugar de una. Una por cada fuente de variación (menos la residual).

- Al factor I: $CMTratI/CMRes = F$ con la hipótesis nula de $H0 : \alpha_i = 0$
- Al factor II: $CMTratII/CMRes = F$ con la hipótesis nula de $H0 : \beta_j = 0$
- A la interacción: $CMInt/CMRes = F$ con la hipótesis nula de $H0 : (\alpha\beta)_{ij} = 0$

En este caso mi **primer interés** es siempre revisar la hipótesis de *interacción*, ya que esto cambia la interpretación.

Es importante recordar que solo puedo dar conclusiones en términos de los tratamientos individualmente si y solo si no hay interacción, ya que si hay el efecto de un nivel de a, depende del nivel de b en el que esté (o como se quiera ver)

Promedios estimados. (Ignorando la interacción)

Estos (prácticamente) nunca van a coincidir con los promedios observados por tratamiento, ya que el efecto de interacción, por más minúsculo o insignificante que sea siempre va a estar ahí (A no ser de que sea literalmente 0), esto es independiente de si se encuentra que la interacción es estadísticamente no significativa.

Por esta razón debemos estimar los promedios con los efectos por nivel del factor.

$$\bar{y}_{ij} = \bar{\bar{y}} + \alpha_i + \beta_j$$

Promedios Observados

Siempre van a contener los efectos de interacción, aunque sean muy mínimos, por eso no coinciden con los estimados. (A no ser de que sea exactamente 0)

Más sobre la interacción en el ANOVA

Es importante siempre considerar la interacción si es significativa. Porque si no estoy inflando el cuadrado medio residual.

Si no rechazo esta hipótesis, es mala idea dejar la interacción en el modelo porque si no estoy gastando grados de libertad.

ANOVA (Ignorando interacción)

En este caso hay que tomar en cuenta varias cosas: - El CMRes ahora contiene la variabilidad que no explican los dos tratamientos, es decir, se incluye la variabilidad que hubiera estado en el cuadrado medio de interacción. Esta aumenta. - Los residuales se calculan ahora respecto a la *media estimada sin interacción*, que es diferente a la media observada, esto causa que aumente la Suma de cuadrados residual. - Es importante recordar que la suma de cuadrados total siempre es fija.

- Recordar que en estos casos se puede hablar independientemente de los efectos de los factores.

Parametrizaciones del modelo

Suma Nula

La lógica es la misma que con un factor pero hay que tener cuidado porque se usa para cada factor y sus diferentes niveles

Necesito a-1 variables auxiliares en el primer factor y b-1 en el segundo.

Ejemplo:

Tratamiento	A_1	A_2	B_1
11	1	0	1
11	1	0	1
21	0	1	1
21	0	1	1
31	-1	-1	1
31	-1	-1	1
12	1	0	-1
12	1	0	-1
22	0	1	-1
22	0	1	-1
32	-1	-1	-1
32	-1	-1	-1

El modelo de suma nula

Sin interacción:

$$E[Y|trat] = \mu + \alpha_1 A_1 + \alpha_2 A_2 + \beta_1 B_1$$

Con interacción:

$$E[Y|trat] = \mu + \alpha_1 A_1 + \alpha_2 A_2 + \beta_1 B_1 + (\alpha\beta)_{11} A_1 B_1 + (\alpha\beta)_{21} A_2 B_1$$

No necesitamos todos los efectos en el modelo porque sabemos que hay equivalencias gracias a las parametrizaciones. Solo es necesario colocar (a-1)(b-1) efectos de interacción.

Entonces la tabla completa quedaría algo así:

Para incluir la interacción debo de multiplicar las variables auxiliares anteriores

Trat	A1	A2	B1	A1B1
11	1	0	1	
21	0	1	1	
31	-1	-1	1	
12	1	0	-1	
22	0	1	-1	
32	-1	-1	-1	

Tratamiento de referencia

Es igual que anteriormente, solo con unos y 0

Trat	A2	A3	B2
11	0	0	0
21	1	0	0
31	0	1	0
12	0	0	1
22	1	0	1
32	0	1	1

Comparaciones múltiples

Recordemos que tenemos una hipótesis importante:

$$H0 : (\alpha\beta)_{ij} = 0$$

Entonces tenemos dos situaciones:

- **CI** Si se rechaza la hipótesis hay interacción quiere decir que no puedo comparar los factores independientemente de cada uno de ellos. En este caso hay que tomar diferentes consideraciones.
  - Se hacen comparaciones de las medias del factor principal para cada nivel del otro factor (se fijan)
  - Se debe de hacer bonferroni cuando es necesario (O cuando se comparan todos los pares con todos)
  - Para hacer bonferroni debo de dividir el alfa entre d, siendo d la cantidad de pares que quiero comparar dentro del nivel del factor que no es principal, es decir si el factor B tiene 2 niveles y el factor A tiene 3 niveles, d sería igual a 3.
    - En ese sentido dentro de los niveles no hay independencia (Ortogonalidad) , entre los niveles si hay independencia.
  - Las hipótesis lucen algo así
    - $H0 : \mu_{11} = \mu_{21}$
  - En este caso se hace fijando en un nivel del factor que no es de diseño para todos los niveles del otro factor.
- **SI** Si no se rechaza la hipótesis: No hay interacción y se quieren comparar todos los pares de un factor se pueden comparar usando Tukey (si el efecto del factor es significativo).
  - En este caso se reajusta el modelo sin tomar en cuenta esa fuente de interacción.
  - Se hacen comparaciones para los factores con *efectos significativos.*
  - Las hipótesis en este caso lucirían algo así:
    - $H0 : \mu_{1.} = \mu_{2.}$
  - Las hipótesis deben de ser con la media marginal debido a un tema de notación, aunque estemos solo en un factor, el otro factor no debe ser ignorado en este aspecto.
  - Las medias y vectores marginales solo pueden ser usados de esta forma con el modelo de suma nula, con tratamiento de referencia es más complejo.
  - Por supuesto, puedo calcular las medias marginales con vectores y el vector de coeficientes del modelo
  - Para calcular la resta de mu1 con mu2, también puedo hacerlo restando los vectores.

EMMEANS

Se pueden hacer automáticamente con la librería emmeans

En este caso, puesto que hay interacción, debe indicarse que se hacen las comparaciones para los promedios de A dentro de cada nivel de B, y se pide que se use la corrección de Boferroni de la siguiente forma: emmeans(mod, pairwise~A|B, adjust="bonferroni")

Las probabilidades de la salida de esta función se deben de interpretar distinto

2alpha si son de una cola y alphas si es de dos colas.

Estas probabilidades no son las mismas que se obtienen a mano.

Si se quieren obtener las mismas probabilidades del punto (e), las probabilidades de la salidad de emmeans deben dividirse por la cantidad de grupos*cantidad de comparaciones, o las obtenidas a mano deben de multiplicarse por este mismo valor.

Ejemplo del modelo sin interacción

En un ejemplo sin interacción, para comparar todos los pares se recurre al método de Tukey, para hacer estas comparaciones debo de recurrir a las medias marginales para cada nivel del factor.

Por ejemplo, usando suma nula, las medias marginales de un factor de 3 niveles se consiguen Dandole un valor de 1 a al nivel de interés, un 0 al que no es de interés y un 0 siempre al factor B.**(este es el marginal)**

Ejemplo:

Para el nivel 1

$$\mu_{1.} = \mu + \alpha_1 * 1 + \alpha_2 * 0 + \beta_1 * 0 = \mu + \alpha_1$$

Para el nivel 2

$$\mu_{1.} = \mu + \alpha_1 * 0 + \alpha_2 * 1 + \beta_1 * 0 = \mu + \alpha_2$$

Para el nivel 3 se deben de hacer A1=-1 y A2=-1

$$\mu_{1.} = \mu + \alpha_1 * -1 + \alpha_2 * -1 + \beta_1 * 0 = \mu + \alpha_3$$

Variables confusoras

Una variable confusora crea un desbalance en el experimento, el cual hace difícil llegar a conclusiones que la variable respuesta sea consecuencia del tratamiento o de otro factor

Por ejemplo:

En el caso de las serpientes, si hay muchas hembras y muy pocos machos en una especie, y hay muchos machos y muy pocas hembras, es posible que haya una confusión, porque no sabriamos si las diferencia de veneno se deben a la especie, o a que los machos de la especie B inyectan una cantidad diferente de veneno que a las hembras de la especie A.

- En este caso el problema de la confusión, el mismo se puede atacar, por ejemplo, registrando el sexo de las serpientes y viendo si hay un desbalance en ese sentido.
- La aleatorización es una forma de atacar la confusión
- En un experimento ya hecho, se puede aumentar el tamaño de muestra siempre y cuando esto no afecte la variable respuesta de alguna forma.
  - Si esto afecta la respuesta es mejor diseñar el experimento desde 0, pero tomando en cuenta el desbalance (es complicado y caro, no siempre se puede hacer)