Diseños Multifactoriales

Clase 7

Nicolás Mejía M.
n.mejia10@uniandes.edu.co

Probabilidad y Estadística II
Departamento de Ingeniería Industrial
Universidad de Los Andes, Bogotá, Colombia

2020-19

Outline

1. Remember, Remember...

2. Diseños Multifactoriales
   - Notación
   - Las sumas de cuadrados
   - Medias Cuadraticas y Grados de Libertad
   - Prueba F
   - Tabla ANOVA

3. Selección de un modelo
   - Principio de Jerarquía

La interacción

Existe la posibilidad que entre los factores exista una efecto adicional que se produce solo al darse de forma simultanea para ciertos tratamientos. Un relación que no se puede aprecia de forma individual.

A ese efecto lo denominamos interacción. Es de interés conocer si esto se produce debido a que su existencia implicaría que nuestros resultados no pueden generalizarse a los niveles de los factores, sino que deben darse a nivel de tratamientos.

De esta forma, se tiene una nueva descomposición de la variabilidad que observemos en los datos:

$$\text{Variación Total} = \text{Var. por factor 1} + \text{Var. por factor 2} + \text{Var. por Interacción} + \text{Var. por efecto aleatorio}$$

La tabla ANOVA

Un diseño de dos factores con interacción se puede organizar en forma de tabla de la siguiente manera:

Fuente	SS	gl	MS	F
A	$\sum_{i=1}^{a} bn\left(\bar{Y}_{i..}-\bar{Y}_{...}\right)^2$	$a-1$	$SSA/(a-1)$	MSA/MSE
B	$\sum_{j=1}^{b} an\left(\bar{Y}_{.j.}-\bar{Y}_{...}\right)^2$	$b-1$	$SSB/(b-1)$	MSB/MSE
AB	$\sum_{i,j}^{a,b} n\left(\bar{Y}_{ij.}-\bar{Y}_{i..}-\bar{Y}_{.j.}+\bar{Y}_{...}\right)^2$	$(a-1)$ $\cdot(b-1)$	$MSAB$	$MSAB/MSE$
Error	$\sum_{i,j}^{a,b}(n-1)S_{ij}^2$	$N-ab$	MSE	
Total	$\sum_{i,j,j}^{a,b,n}\left(Y_{ijk}-\bar{Y}_{...}\right)^2$	$N-1$	$SST/(N-1)$	

Donde $SST = SSA + SSB + SSAB + SSE$ y $gl_T = gl_A + gl_B + gl_{AB} + gl_E$

Outline

Diseños con Múltiples Factores

La idea, es usar las mismas unidades experimentales (sobre las cuales se mide la variable de respuesta), para concluir sobre la influencia simultánea de varios factores.

Ejemplos:

- Cuando la variable de respuesta son la ventas de pan, un factor puede ser la posición vertical en el estante (arriba, centro o abajo); el otro la posición horizontal (izquierda o derecha) y otro el precio (alto o bajo).
- Cuando se quiere investigar sobre factores que afectan el nivel de colesterol, se puede usar el consumo de huevo (si o no), el nivel de actividad física que se realiza (alta, meda o baja) y la calidad del sueño (buena o mala).

La idea es que al diseñar un experimento, se escojan los factores de interés que se quieren investigar.

Diseños Multifactoriales

Vamos a ver el caso general en el que hay muchos factores.

El factor A tendrá a niveles, el factor B tendrá b niveles, el factor C tendrá c niveles, el factor D tendrá d niveles, etc.

Un tratamiento es una combinación de niveles de los factores (i.e un tratamiento corresponde a la tupla del i-ésimo nivel del factor 1, j-ésimo nivel del factor 2, k-ésimo nivel del factor 3, l-ésimo nivel del factor 4, etc).

El número de réplicas, es decir, cantidad de veces que se realiza el experimento bajo un tratamiento dado, se denotará n, donde asumiremos que el diseño es balanceado, a menos que se diga lo contrario.

Diseños Multifactoriales

La notación de las variables que producen los datos y la representación tabular de los datos se complica un poco con el número de factores. Por ejemplo:

Para el caso de tres factores

Y_{ijkl} representa la l-ésima observación de la variable de interés correspondiente al i-ésimo nivel del factor A, el j-ésimo nivel del factor B y el k-ésimo nivel del factor C.

$$Y_{ijkl} \sim Normal\left(\mu_{ijk}, \sigma^2\right)$$

En general, si se tienen más de 3 factores, ya no se usa este tipo de notación explícita, pero los conceptos se mantienen igual.

Outline

Notación

Promedios por nivel

$$\bar{Y}_{i\ldots} = \frac{1}{bcn}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{n} Y_{ijkl}$$

$$\bar{Y}_{.j.} = \frac{1}{acn}\sum_{i=1}^{a}\sum_{k=1}^{c}\sum_{l=1}^{n} Y_{ijkl}$$

$$\bar{Y}_{..k.} = \frac{1}{abn}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{l=1}^{n} Y_{ijkl}$$

Promedio por tratamiento

$$\bar{Y}_{ijk.} = \frac{1}{n}\sum_{l=1}^{n} Y_{ijkl}$$

Promedio total

$$\bar{Y}_{....} = \frac{1}{N}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{n} Y_{ijkl}$$

donde $N = abcn$, el total de los datos.

Otros promedios

$$\bar{Y}_{ij..} = \frac{1}{cn}\sum_{k=1}^{c}\sum_{l=1}^{n} Y_{ijkl}$$

$$\bar{Y}_{.jk.} = \frac{1}{an}\sum_{i=1}^{a}\sum_{l=1}^{n} Y_{ijkl}$$

$$\bar{Y}_{i.k.} = \frac{1}{bn}\sum_{j=1}^{b}\sum_{l=1}^{n} Y_{ijkl}$$

Tabla

El diseño realmente es un cubo. Una de las caras sería:

UNA CARA DEL CUBO		Factor A					
		Nivel 1	...	Nivel i	...	Nivel a	Prom. Fila
Factor B	Nivel 1	$\bar{Y}_{11..}$ $S_{11.}$	... ...	$\bar{Y}_{i1..}$ $S_{i1.}$	... ...	$\bar{Y}_{a1..}$ $S_{a1.}$	$\bar{Y}_{.1..}$
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	Nivel j	$\bar{Y}_{1j..}$ $S_{1j.}$	... ...	$\bar{Y}_{ij..}$ $S_{ij.}$	... ...	$\bar{Y}_{aj..}$ $S_{aj.}$	$\bar{Y}_{.j..}$
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	Nivel b	$\bar{Y}_{1b..}$ $S_{1b.}$	... ...	$\bar{Y}_{ib..}$ $S_{ib.}$	... ...	$\bar{Y}_{ab..}$ $S_{ab.}$	$\bar{Y}_{.b..}$
	Prom. columna	$\bar{Y}_{1...}$	...	$\bar{Y}_{i...}$	...	$\bar{Y}_{a...}$	$\bar{Y}_{....}$

Outline

Las sumas de cuadrados

Las sumas de cuadrados cuando se tienen más factores se definen igual.
La expresión cambio visualmente un poco debido a los índices, PERO SON LAS MISMAS FORMULAS DE ANTES.

Suma de cuadrados TOTAL

$$SST = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{n}\left(Y_{ijkl} - \bar{Y}_{....}\right)^2$$

Suma de cuadrados de un Factor

$$SSA = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{n}\left(\bar{Y}_{i...} - \bar{Y}_{....}\right)^2 = \sum_{i=1}^{a} bnc\left(\bar{Y}_{i...} - \bar{Y}_{....}\right)^2$$

Se usa la misma lógica para los demas factores.

Las sumas de cuadrados

Suma de cuadrados de una interacción DOBLE

$$SSAB = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{n}\left(\bar{Y}_{ij..} - \bar{Y}_{i...} - \bar{Y}_{.j..} + \bar{Y}_{....}\right)^2$$

$$= \sum_{i=1}^{a}\sum_{j=1}^{b} cn\left(\bar{Y}_{ij..} - \bar{Y}_{i...} - \bar{Y}_{.j..} + \bar{Y}_{....}\right)^2$$

La sumas de cuadrados de interacciones de mayor orden

También puede darse un efecto especial cuando 3 o más factores afectan en simultaneo. Luego pueden existir interacciones triples, cuadruples, etc.

Las sumas de cuadrados

La suma de cuadrados del error

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{n}\left(Y_{ijkl} - \bar{Y}_{ijk\cdot}\right)^2$$

Ecuación Fundamental

$$\begin{aligned} SST &= SSA + SSB + SSC && \text{Efectos Individuales} \\ &+ SSAB + SSAC + SSCB && \text{Interacciones Dobles} \\ &+ SSABC && \text{Interacción Triple} \\ &+ SSE && \text{Error} \end{aligned}$$

Y así sucesivamente cuando se tienen mas factores.

Outline

Medias Cuadraticas y Grados de Libertad

En general se definen las medias cuadraticas como $MSX = \frac{SSX}{gl_X}$, donde:

Grados de libertad totales

$$gl_T = N - 1$$

Grados de libertad de los factores

$$gl_A = a - 1$$

$$gl_B = b - 1$$

$$gl_C = c - 1$$

etc

Número de niveles menos 1

Medias Cuadraticas y Grados de Libertad

Grados de libertad de las interacciones dobles

$$gl_{AB} = (a-1)(b-1)$$

$$gl_{AC} = (a-1)(c-1)$$

$$gl_{BC} = (b-1)(c-1)$$

etc

Se multiplican los grados de libertad individuales

Grados de libertad de las interacciones de orden mayor

$$gl_{ABC} = (a-1)(b-1)(c-1)$$

Misma Lógica, Se multiplican los grados de libertad individuales

Grados de libertad

Ecuación Fundamental

$$
\begin{aligned}
gl_T = gl_A + gl_B + gl_C & \quad \text{Efectos Individuales} \\
+gl_{AB} + gl_{AC} + gl_{BC} & \quad \text{Interacciones Dobles} \\
+gl_{ABC} & \quad \text{Interacción Triple} \\
+gl_E & \quad \text{Error}
\end{aligned}
$$

Grados de libertad del Error

Ya conocemos como calcular los otros, luego los grados de libertad del error salen a cuadre, despejando de la ecuación fundamental.

Outline

Prueba F

Interacción X

$$H_0: \text{ La Interacción X NO influye sobre } Y$$
$$H_1: \text{ La Interacción X SI influye sobre } Y$$

Factor X

$$H_0: \text{ El Factor X NO influye sobre } Y \Leftrightarrow \mu_{1.} = \mu_{2.} = \cdots = \mu_{a.}$$
$$H_1: \text{ El Factor X SI influye sobre } Y \Leftrightarrow \text{Algún par } \mu_{i.} \neq \mu_{j.}$$

Las pruebas se hacen en orden de complejidad: Primeros las triples, luego las dobles, luegos si las individuales.

Prueba F

Bajo la hipótesis nula, el estadístico dado por:

Estadístico

$$F = \frac{\frac{\frac{SSX}{\sigma^2}}{gl_X}}{\frac{\frac{SSE}{\sigma^2}}{gl_E}} = \frac{\frac{SSX}{gl_X}}{\frac{SSE}{gl_E}} = \frac{MSX}{MSE} \sim F_{gl_X 1, gl_E}$$

Valores grandes del estadístico F están a favor de que el factor/interaccion SI es significativo, mientras que valores pequeños son evidencia NO son significativo. Estadísticamente tenemos la siguiente región de rechazo: $RHO \Leftrightarrow F \geq F_{1-\alpha, gl_X, gl_E}$

Outline

Tabla ANOVA

En formato de tabla queda para 3 factores:

Fuente	**SS**	**gl**	**MS**	**F**
Factor A	SSA	$a-1$	MSA	$F = \frac{MSA}{MSE}$
Factor B	SSB	$b-1$	MSB	$F = \frac{MSB}{MSE}$
Factor B	SSC	$c-1$	MSB	$F = \frac{MSC}{MSE}$
Interac. AB	SSAB	$(a-1)(b-1)$	MSAB	$F = \frac{MSAB}{MSE}$
Interac. AC	SSAC	$(a-1)(c-1)$	MSAC	$F = \frac{MSAC}{MSE}$
Interac. BC	SSBC	$(c-1)(b-1)$	MSBC	$F = \frac{MSBC}{MSE}$
Interac. ABC	SSABC	$(a-1)(b-1)(c-1)$	MSABC	$F = \frac{MSABC}{MSE}$
Error	SSE	$abc(n-1)$	MSE	
Total	SST	$N-1$		

Outline

Idea

Una vez se tiene el diseño factorial, ¿Qué sucede si una fuente de variación es NO significativa? Si realmente NO afecta, entonces ¿por qué considerarla en el diseño? La intuición nos dice que la quitemos.

Al quitar una fuente de variación del ANOVA, su suma de cuadrados se suma a la del error (lo que no explicamos). No es que perdamos una fuente de explicación, pues realmente nunca lo fue. Realmente es una parte del error que se atribuyó erróneamente a dicha posible fuente de variación.

Ya mencionamos que el MSE será el mejor estimador de la varianza y que los grados de libertad son la cantidad efectiva de términos (recursos) utilizados en esa estimación. Al enviar una fuente de variación al error, también se le suman los grados de libertad. Luego, es mejor estimar la varianza con más recursos! luego más grados de libertad en el error son bienvenidos.

Idea

¿Qué es eso de mandar al error? Cuando íbamos agregando filas a la tabla ANOVA, decíamos que el error actual se descomponía en las nuevas sumas de cuadrados que se agregaban, más un nuevo error.

$$\text{ANOVA 1 factor } SST = SSA + \widetilde{SSE}$$

$$\text{ANOVA 2 factores con interacción } SST = SSA + SSB + SSAB + SSE$$

$$\text{Luego } \widetilde{\text{SSE}} = \text{SSB} + \text{SSAB} + \text{SSE}$$

Mandar al error es hacer el procedimiento inverso, es decir, recomponer el error al agrupar sumas de cuadrados.

Outline

Principio de Jerarquía

¿En que orden enviamos las sumas de cuadrados al error?

SI hay un orden que se debería seguir, para evitar ciertas incongruencias, más que todo conceptuales.

Ejemplo

Suponga un caso en que la interacción entre dos factores da significativa, pero los factores individuales no son significativos. ¿Tiene sentido quitar los factores? NO! Sucede todo lo contrario. El efecto individual existe, pero se ve reflejado solo en la interacción. Luego, no tiene sentido quitar las sumas de cuadrados individuales. No tiene sentido tener una interacción doble, si no se tiene los factores individuales.

Este ejemplo nos da una idea de como empezar a mirar que sumas de cuadrados enviamos al error, según su significancia.

Principio de Jerarquía

Principio de Jerarquía

En palabras simples sugiere que se analice lo mas complejo primero (interacciones de orden superior) y se evalúe su significancia. Si es significativo, dejar el modelo como esta, si no es significativo, quitarla y luego evaluar lo siguiente mas complejo.

En ese orden de ideas, por ejemplo en un ANOVA de 3 factores con interacciones, primero se miraría la triple interacción, luego las interacciones dobles y finalmente los factores individuales.

De esta forma, el ANOVA se simplifica y solo quedan los efectos que sean estadisticamente significativos, al tiempo que se tienen conceptualmente con sentido.

Veamos un ejemplo específico

Principio de jerarquía: Combinación Efectos de Variabilidad

Retomando todo lo visto de la prueba ANOVA y de los experimentos con dos factores, sabemos que:

Efecto	Si H_0 es cierto	Si H_1 es cierto
Error	$E(MSE) = \sigma^2$	$E(MSE) = \sigma^2$
Factor A	$E(MSA) = \sigma^2$	$E(MSA) > \sigma^2$
Factor B	$E(MSB) = \sigma^2$	$E(MSB) > \sigma^2$
Interacción AB	$E(MSAB) = \sigma^2$	$E(MSAB) > \sigma^2$

Principio de jerarquía: Combinación Efectos de Variabilidad

Ahora suponga que para alguno de los efectos ($X = A,\ B,\ AB$), la hipótesis nula es cierta, es decir, este efecto no es significativo. En este caso $E(MSX) = \sigma^2$.

¿Porqué no combinar el efecto X con el error para estimar la varianza σ^2 con más grados de libertad (esto es, con un tamaño efectivo de muestra mayor)?

Aditividad de las Sumas de Cuadrados

- Esta combinación si es posible, la manera de hacerlo, sin embargo, es buscando que el estimador resultante para σ^2 sea el mejor (insesgado con menor varianza).
- El resultado, es que si se combinan las medias cuadráticas de los efectos no significativos, ponderada por los respectivos grados de libertad, se obtiene un buen estimador .

Principio de jerarquía: Aditividad de las Sumas de Cuadrados

Como un caso particular, suponga que se sabe que la interacción AB no tiene efecto. En este caso,

$$E(MSAB) = E\left(\frac{SSAB}{(a-1)(b-1)}\right) = \sigma^2 \quad \Rightarrow$$

$$MSE^{new} = \frac{(a-1)(b-1)MSAB + ab(n-1)MSE}{(a-1)(b-1) + ab(n-1)} = \frac{SSAB + SSE}{N - a - b + 1} \quad \Rightarrow$$

$$E\left(MSE^{new}\right) = \sigma^2 \quad \Rightarrow$$

$$\frac{SSE^{new}}{\sigma^2} \sim \chi^2_{N-a-b+1}$$

Nota: Este MSE^{new} coincide con el que se definió en un diseño con dos factores. ¿Porqué?