ANOVA de 2 Factores con Interacción

Clase 6

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

Idea

Ya sabemos que:

$$\text{Variación Total} = \text{Var. por efecto de grupos} + \text{Var. por efecto aleatorio}$$

Donde el término del error hace alusión a toda la variación presente en los datos que el factor no puede explicar.

Esta variación simplemente proviene de otras fuentes que no están siendo consideradas en el diseño actual del ANOVA.

En otras palabras, hay otros FACTORES que pueden estar influenciando nuestra variable de interés.

La idea de la clase de hoy es discutir como incluirlos en el ANOVA

Notación

Cuando se tienen dos factores Y_{ijk} representa la variable de interés para la k-ésima observación del i-ésimo nivel del factor 1 y el j-ésimo nivel del factor 2. De nuevo, se asume que $Y_{ijk} \sim Normal(\mu_{ij}, \sigma^2)$.

Ahora los promedios son:

$$\bar{Y}_{i..} = \frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n} Y_{ijk}, i \in \{1,\ldots,a\} \quad \bar{Y}_{.j.} = \frac{1}{an}\sum_{i=1}^{a}\sum_{k=1}^{n} Y_{ijk}, j \in \{1,\ldots,b\}$$

$$\bar{Y}_{ij.} = \frac{1}{n}\sum_{k=1}^{n} Y_{ijk}\ i \in \{1,\ldots,a\}\ j \in \{1,\ldots,b\}$$

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

Con $N = abn$, siendo este valor la totalidad de datos.

Notación

Al organizar esta información en una tabla, queda:

ORDEN DE DATOS		Factor A					
		Nivel 1	...	Nivel i	...	Nivel a	Prom. Fila
Factor B	Nivel 1	Y_{111}, Y_{112} $\ldots, Y_{11n}$	$\ldots$ $\ldots$	Y_{i11}, Y_{i12} $\ldots, Y_{i1n}$	$\ldots$ $\ldots$	Y_{a11}, Y_{a12} $\ldots, Y_{a1n}$	$\bar{Y}_{.1.}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	Nivel j	Y_{1j1}, Y_{1j2} $\ldots, Y_{1jn}$	$\ldots$ $\ldots$	Y_{ij1}, Y_{ij2} $\ldots, Y_{ijn}$	$\ldots$ $\ldots$	Y_{aj1}, Y_{aj2} $\ldots, Y_{ajn}$	$\bar{Y}_{.j.}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	Nivel b	Y_{1b1}, Y_{1b2} $\ldots, Y_{1bn}$	$\ldots$ $\ldots$	Y_{ib1}, Y_{ib2} $\ldots, Y_{ibn}$	$\ldots$ $\ldots$	Y_{ab1}, Y_{ab2} $\ldots, Y_{abn}$	$\bar{Y}_{.b.}$
	Prom. columna	$\bar{Y}_{1..}$	$\ldots$	$\bar{Y}_{i..}$	$\ldots$	$\bar{Y}_{a..}$	$\bar{Y}_{...}$

Notación

Con los promedios y desviaciones por tratamiento, la tabla queda:

RESUMEN DE DATOS		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

Motivación

Cuando consideramos un diseño con 2 factores, en principio podríamos hacer 2 experimentos separados. ¿Qué nos estamos perdiendo?

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.

Ejemplo

Suponga que en un experimento está considerando como factor 1 una bebida gaseosa y como factor 2 ciertas referencias de mentas. Individualmente pueden que no haya un efecto particular en una persona, pero cuando se aplican al mismo tiempo...

La interacción

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:

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

Outline

La Suma de Cuadrados

Ya habíamos definido las sumas de cuadrados total y de os factores. recuerden que estas son iguales independiente del diseño experimental.

Suma de cuadrados total

$$SST = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \bar{Y}_{...})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} Y_{ijk}^2 - N\bar{Y}_{...}^2$$

Suma de cuadrados de los factores A y B

$$SSA = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{Y}_{i..} - \bar{Y}_{...})^2 = \sum_{i=1}^{a} bn(\bar{Y}_{i..} - \bar{Y}_{...})^2$$

$$SSB = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{Y}_{.j.} - \bar{Y}_{...})^2 = \sum_{j=1}^{b} an(\bar{Y}_{.j.} - \bar{Y}_{...})^2$$

Suma de cuadrados de la interacción

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

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

Intento de Explicación

Haciendo álgebra en la expresión anterior, el término sin el cuadrado en la suma se puede escribir como: $\bar{Y}_{ij.} - (\bar{Y}_{...} + (\bar{Y}_{i..} - \bar{Y}_{...}) + (\bar{Y}_{.j.} - \bar{Y}_{...}))$. Ya vimos que $(\bar{Y}_{i..} - \bar{Y}_{...})$ y $(\bar{Y}_{.j.} - \bar{Y}_{...}))$ corresponden al efecto de los factores, luego $(\bar{Y}_{...} + (\bar{Y}_{i..} - \bar{Y}_{...}) + (\bar{Y}_{.j.} - \bar{Y}_{...}))$ es un promedio corregido que incluye dichos efectos. Por tanto, $\bar{Y}_{ij.} - (\bar{Y}_{...} + (\bar{Y}_{i..} - \bar{Y}_{...}) + (\bar{Y}_{.j.} - \bar{Y}_{...}))$ es la diferencia entre el promedio del tratamiento y los efectos individuales de los factores, es decir, un efecto adicional (interacción) que no contemplan los factores marginalmente.

La Suma de Cuadrados del Error

Finalmente el error hace alusión a lo que no es explicado por medio de los factores y su interacción. En este caso la suma de cuadrados del error, sería las desviaciones que se producen de un dato con respecto a su promedio por tratamiento::

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \bar{Y}_{ij.})^2$$

Nuevamente la expresión del SSE cambia brutalmente con la que ya se tenía.

La Ecuación Fundamental

Igualmente:

La Ecuación Fundamental de las Sumas de Cuadrados

$$SST = SSA + SSB + SSAB + SSE$$

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

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

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

Outline

Grados de libertad

Ya sabemos

Grados de libertad total y de los factores

$$SST = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \bar{Y}_{...})^2 \Rightarrow gl_T = N - 1$$

$$SSA = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{Y}_{i..} - \bar{Y}_{...})^2 \Rightarrow gl_A = a - 1$$

$$SSB = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{Y}_{.j.} - \bar{Y}_{...})^2 \Rightarrow gl_B = b - 1$$

Grados de libertad

Grados de libertad de la Interacción

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

Cuantos $\bar{Y}_{ij.}$ hay? Hay ab tratamientos. ¿Cuantos $\bar{Y}_{i..}$ hay? Hay a y van restando. ¿Cuantos $\bar{Y}_{.j.}$ hay? Hay b y van restando. ¿Cuantos $\bar{Y}_{...}$ hay? Hay solo uno. Agregando se tienen $ab - a - b + 1 = (a-1)(b-1)$ grados de libertad del SSAB.

Grados de libertad del error

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \bar{Y}_{ij.})^2$$

¿Cuantos Y_{ijk} hay? Hay N datos. ¿Cuantos $Y_{ij.}$ hay? Hay ab y van restando. Agregando se tienen $N - ab$ grados de libertad del SSE.

Outline

Medias Cuadráticas

Igual que antes se definen las medias cuadráticas:

$$MST = \frac{SST}{gl_T} = \frac{SST}{N-1} \qquad MSA = \frac{SSA}{gl_A} = \frac{SSA}{a-1}$$

$$MSB = \frac{SSB}{gl_B} = \frac{SSB}{b-1} \qquad MSAB = \frac{SSAB}{gl_{AB}} = \frac{SSAB}{ab-a-b+1}$$

$$MSE = \frac{SSE}{gl_E} = \frac{SSE}{N-ab}$$

Igual que en el caso de 1 factor, la ecuación fundamental también aplica para los grados de libertad:

$$gl_T = gl_A + gl_B + gl_{AB} + gl_E$$

$$N-1 = (a-1) + (b-1) + (ab-a-b+1) + (N-ab)$$

Una vez más lo que cambian son los grados de libertad del error.

Outline

Prueba F

Las preguntas de interes son:

Interacción

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

Factor A

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

Factor B

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

Las probaremos en ese orden.

Prueba F

Bajo la validez de la hipótesis nula, donde los factores e interacción NO son significativos, se cumple lo siguiente:

Aplicación del teorema de Cochran

Las sumas de cuadrados anteriores, divididas por la varianza σ^2, se distribuyen χ^2 con sus respectivos grados de libertad:

$$\frac{SST}{\sigma^2} = \frac{SSA}{\sigma^2} + \frac{SSB}{\sigma^2} + \frac{SSAB}{\sigma^2} + \frac{SSE}{\sigma^2}$$

$$\chi^2_{N-1} = \chi^2_{a-1} + \chi^2_{b-1} + \chi^2_{ab-a-b+1} + \chi^2_{N-ab}$$

De igual manera que en el ANOVA de un factor, vamos a ver si el SSA, el SSB y el SSAB, son estadísticamente "grandes"con respecto al SSE.

Prueba F

Bajo la hipótesis nula, el estadístico dado por:

Interacción

$$F = \frac{\frac{\frac{SSAB}{\sigma^2}}{ab-a-b+1}}{\frac{\frac{SSE}{\sigma^2}}{N-a-b+1}} = \frac{\frac{SSAB}{ab-a-b+1}}{\frac{SSE}{N-ab}} = \frac{MSAB}{MSE} \sim F_{ab-a-b+1,N-ab}$$

Factores

$$F = \frac{MSA}{MSE} \sim F_{a-1,N-ab} \qquad F = \frac{MSB}{MSE} \sim F_{b-1,N-ab}$$

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: $RHO \Leftrightarrow F \geq F_{1-\alpha,gl_X,N-ab}$

Outline

La tabla ANOVA

Toda esta información 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

Caso 1 datos por tratamiento

En las cuentas que desarrollamos, que sucede si solo existiese 1 dato por tratamiento (i.e $n = 1$)

¿Que sucede con el SSE? Da 0, luego no puedo calcular el MSE y los estadísticos F.

No es que estemos explicando perfectamente la variabilidad. Así como lo sugieren los grados de libertad del error en ese caso ($gl_E = N - ab = ab - ab = 0$), nos quedamos sin un número efectivo de datos suficiente para hacer la estimación!

¿Que hacer?

En ese caso, se asume que no hay interacción y se trabaja como el ANOVA de 2 factores sin interacción de la clase pasada. No es que no haya interacción, pero simplemente no la podemos medir.

Ejemplo

La universidad de Chicago quiere estudiar dos posibles efectos sobre el desempeno de sus estudiantes en el Examen GMAT. El primero de estos efectos es la cantidad de horas de preparación que tuvo cada estudiantes, distrinuidos así: (I) un repaso de 3 horas, (II) un curso preparatorio de un día y (III) Un curso de una semana. El otro posible efecto se debe a que este examen es presentado usualmente por estudiantes de 3 facultades diferentes: Administración, Ingeniería y Matemáticas. Para esto se ha tomado una muestra aleatoria de los puntajes de dos estudiantes para cada tratamiento. A continuación, se muestan el promedio y la varianza de cada tratamiento:

Ejemplo

	Horas de Preparación		
	I	**II**	**III**
Matemáticas	4.95	4.65	1.95
	0.07	0.35	0.08
Ingeniería	3.95	4.85	1.55
	0.07	0.21	0.64
Administración	3.85	2.50	1.00
	0.07	0.28	0.14

Realice un ANOVA de dos factores con interacción para encontrar los efectos sobre el desempeño de los alumnos de la universidad en el GMAT.