ANOVA de 2 Factores: Bloques

Clase 5

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

Idea

Suponga que ahora tenemos 2 factores influenciado sobre nuestra variable respuesta.

Ejemplo

Considere la estatura de una persona seleccionada de forma aleatoria. La estatura promedio es diferente si hacemos distinción entre hombres y mujeres (FACTOR 1), también será diferente si hacemos distinción entre grupos de edades (FACTOR 2).

Luego, usando la misma lógica que llevamos, la variación total de los datos se puede descomponer en varias fuentes:

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

Pero cuidado, ¡Este último componente aleatorio no es el mismo de antes!

Outline

Notación

Asumiremos que el factor 1 tiene a niveles y que el factor 2 tiene b niveles.

Un tratamiento es una combinaciones de niveles de ambos factores (i.e un tratamiento corresponde a la pareja i-ésimo nivel del factor 1 junto al j-ésimo nivel del factor 2).

El número de réplicas, es decir, cantidad de veces que se realiza el experimento bajo un tratamiento dado se denota n_{ij}. Asumiremos que el diseño experimental es balanceado, luego el número de réplicas no cambia entre tratamientos (i.e $n_{ij} = n$)

Diseños Balanceados

De ahora en adelante, y para experimentos multifactoriales, los diseños serán balanceados, a menos que se diga lo contrario. Este es un supuesto importante para definir las sumas de cuadrados respectivas.

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	$\bar{Y}_{1..}$	...	$\bar{Y}_{i..}$	...	$\bar{Y}_{a..}$	$\bar{Y}_{...}$

Outline

La Suma de Cuadrados Total

La suma de cuadrados TOTAL que ya teníamos definida la vez pasada es en la notación de 2 factores la siguiente expresión:

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

Aquí se mide TODA la variación de un dato con respecto a su comportamiento común dado por el promedio, sin hacer distinción por algún factor.

La suma de cuadrados TOTAL no depende del diseño factorial que se este haciendo, solo es la variación de los datos. Luego, dada una muestra de datos, este será UN VALOR FIJO.

La Suma de Cuadrados del Factor A

Si el factor A NO es significativo, entonces los promedios por niveles serían iguales entre ellos, y en particular al promedio total (i.e. $\bar{Y}_{i..} \approx \bar{Y}_{...}$). De ahí definimos la suma de cuadrados del FACTOR A, que en la nueva notación es:

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

La suma de cuadrados del FACTOR no depende del diseño factorial que se este haciendo. El efecto de un factor es independiente de los efectos de los otros. Luego, dada una muestra de datos, este será UN VALOR FIJO.

La Suma de Cuadrados del Factor B

Ahora que hay otro factor, cómo se medirá su efecto? Pues exactamente igual

La resta $(\bar{Y}_{.j.} - \bar{Y}_{...})$ es una medida de discrepancia entre la homogeneidad de las medias, luego la suma de cuadrados del FACTOR B es:

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

La suma de cuadrados del FACTOR no depende del diseño factorial que se este haciendo. El efecto de un factor es independiente de los efectos de los otros. Luego, dada una muestra de datos, este será UN VALOR FIJO.

La Suma de Cuadrados del Error

La variación dada por el efecto aleatorio hace alusión a lo que no es explicado por medio de los factores, por ese motivo se define la suma de cuadrados del ERROR:

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

Mucho cuidado! La expresión del SSE cambia brutalmente con la que ya se tenía.

La Suma de Cuadrados del Error

Intento de Explicación

Haciendo álgebra en la expresión anterior, el término sin el cuadrado en la suma se puede escribir como: $Y_{ijk} - (\bar{Y}_{...} + (\bar{Y}_{i..} - \bar{Y}_{...}) + (\bar{Y}_{.j.} - \bar{Y}_{...}))$. 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, $Y_{ijk} - (\bar{Y}_{...} + (\bar{Y}_{i..} - \bar{Y}_{...}) + (\bar{Y}_{.j.} - \bar{Y}_{...}))$ es la diferencia entre el dato y el los efectos de los factores, es decir, lo que finalmente no se pudo explicar!

La Ecuación Fundamental

De esta forma, tal como se describió intuitivamente y ya se tenía de antes, se puede demostrar matemáticamente la siguiente relación:

La Ecuación Fundamental de las Sumas de Cuadrados

$$SST = SSA + SSB + SSE$$

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

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

La Suma de Cuadrados del Error

Si comparamos con respecto al ANOVA de un factor donde $SST = SSA + \widetilde{SSE}$, donde se tiene que el $\widetilde{SSE}$ se reparte en las nuevas sumas de cuadrados que acabamos de definir para el otro factor y el nuevo error. Es decir

$$\widetilde{SSE} = SSB + SSE$$

Es decir, la información sale de lo que originalmente no se podía explicar, y con eso estamos reduciendo el error. Esto mejorará la calidad de nuestras conclusiones.

Esto sugiere que a futuro, los nuevos efectos que se vayan definiendo se podrán extraer del error.

Outline

Medias Cuadráticas

Las sumas de cuadrados están sumando sobre todos los datos, luego si queremos ver en promedio como es esa variabilidad, debemos dividir por la cantidad de términos efectivamente utilizados.

Los datos son como un recurso que se va gastando a medida que hacemos estimaciones, luego al hacer una cuenta (como las sumas de cuadrados), realmente no se están utilizando todos los datos, sino los que van quedando.

Hay una forma informal para hallar los grados de libertad: Simplemente se suman la cantidad de términos positivos y se restan los negativos.

Simbólicamente, los términos en negativo hacen alusión a los " recursos que se van gastando".

Grados de libertad

Grados de libertad totales

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

¿Cuantos Y_{ijk} hay? Hay N datos. ¿Cuantos $\bar{Y}_{...}$ hay? Solo uno y va restando. Agregando se tienen $N - 1$ grados de libertad del SST.

Grados de libertad del factor A

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

¿Cuantos $\bar{Y}_{i..}$ hay? Hay a. ¿Cuantos $\bar{Y}_{...}$ hay? Solo uno y va restando. Agregando se tienen $a - 1$ grados de libertad del SSA.

Grados de libertad

Grados de libertad del factor B

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

¿Cuantos $\bar{Y}_{.j.}$ hay? Hay b. ¿Cuantos $\bar{Y}_{...}$ hay? Solo uno y va restando. Agregando se tienen $b-1$ grados de libertad del SSB.

Grados de libertad del error

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

¿Cuantos Y_{ijk} hay? Hay N datos. ¿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 $N - a - b + 1$ grados de libertad del SSE.

Outline

Medias Cuadráticas

Igual que antes se definen las medias cuadráticas:

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

$$MSB = \frac{SSB}{gl_B} = \frac{SSB}{b-1} \qquad MSE = \frac{SSE}{gl_E} = \frac{SSE}{N-a-b+1}$$

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

$$gl - T = gl_A + gl_B + gl_E$$

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

Lo que cambian son los grados de libertad del error.

Outline

Prueba F

Las preguntas de interes son:

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

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

Construiremos un estadístico F para cada una de las hipótesis de la misma manera que lo hicimos en el ANOVA de un factor.

Prueba F

Bajo la validez de la hipótesis nula, donde los factores 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:

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

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

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

Prueba F

Bajo la hipótesis nula, el estadístico dado por:

Factor A

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

Factor B

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

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

Outline

Cambios

¿Qué cambia con respecto al ANOVA de 1 factor?

Lo que cambia principalmente es el ERROR. Tanto su magnitud, como sus grados de libertad. Eso hace que estadísticamente puedan cambiar las conclusiones.

Si bien, hacer ANOVAs de un factor por separado es posible, es mejor tener un diseño que incluya todo de forma simultanea. El error tiene incoporada toda la información de ambos experimentos.

Outline

La tabla ANOVA

Toda esta información se puede organizar en forma de tabla de la siguiente manera:

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

Donde $SST = SSA + SSB + SSE$ y $gl_T = gl_A + gl_B + gl_e$

Outline

Diseño por bloques

Un caso particular del ANOVA de 2 factores es el diseño por bloques.

En algunos casos se está interesado en un factor en especifico, pero la población no es uniforme y existen diferencias substanciales entre los individuos.

Ejemplo

Usted está interesado en el tiempo en que se demora de ir desde de su casa hasta la universidad. Para eso diseña un experimento en el que utiliza diferentes rutas, y pretende ver cuál es la mejor. A la hora de recolectar los datos, hay ocasiones en las que llovió y otras en las que hizo sol.

En ese caso, los datos no son uniformes y estarán claramente diferenciados según clima. Esa influencia no es de interés, pero si puede perturbar la calidad del experimento.

Diseño por bloques

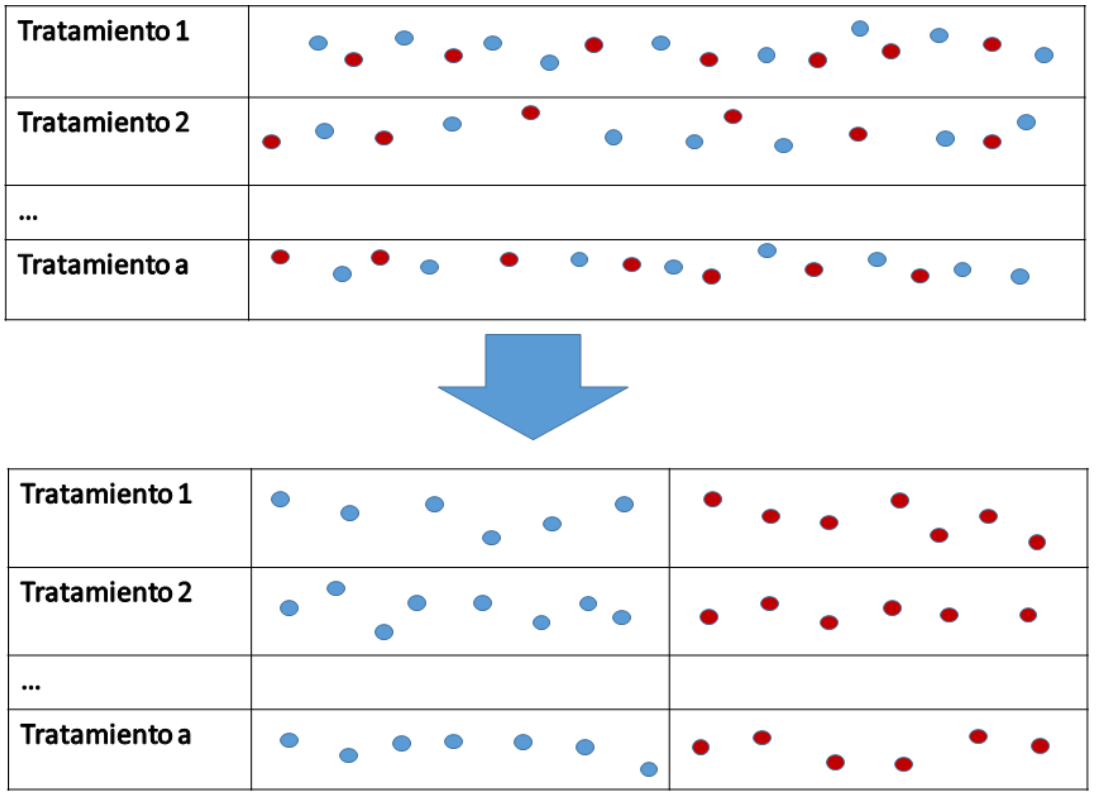
La eficiencia de la prueba para determinar si un factor influye o no sobre la variable de respuesta (Y) depende de qué tan variables son los datos dentro de cada muestra (MSE).

Si la variabilidad es muy grande, es difícil encontrar diferencias significativas entre las medias.

A veces, agrupando las unidades en bloques homogéneos dentro de cada tratamiento ayuda a disminuir el estimador de la varianza aumentando la POTENCIA de la prueba.

Diseño por bloques

¿De qué se trata?: al estimar σ^2 con el *MSE*, se tiene en cuenta que este estimador está INFLADO por la posible diferencia entre los bloques. Esta variabilidad extra se puede cuantificar y sacar del estimador:



Diseño por bloques

Esta formulación es equivalente a un ANOVA de 2 factores., En lugar de tener SSB, tendríamos $SSBLOQUE$, pero matemáticamente hablando se utiliza el mismo proceso que vimos.

La única diferencia es conceptual, donde el segundo factor es el bloque, el cual, solo se incluye para eliminar la variabilidad y poder concluir más acertadamente con respecto al factor que si es de interés.

Ejemplo

En el ejemplo de la ruta, podría incluir el estado del tiempo como un segundo factor (lluvia, no lluvia) que vendría siendo el bloque. La significancia o no del estado del tiempo no interesa (de hecho ya sabemos que es significativo), el propósito es quitar ese ruido para elegir la mejor ruta.

Los bloques se utilizan cuando ese factor de ruido puede ser muy costoso de controlar.

Ejemplo

Suponga que usted es un inversionista y ha consultado a diferentes bancos de inversión con el fin de que le ayuden a estimar el retorno esperado mensual de cuatro portafolios posibles para invertir. Se seleccionó de manera aleatoria un grupo de analistas de cada uno de los bancos, los cuales presentaron una estimación de acuerdo a tres estados diferentes de la economía. Los portafolios se componen de acciones de Ecopetrol, Davivienda y Avianca, en proporciones distintas para cada portafolio. Tenga en cuenta que los portafolios son independientes. La información entregada por los analistas se resume en la siguiente tabla:

	Escenario		
	Pesimista	**Neutral**	**Optimista**
Portafolio 1	-2.5	-0.21	4.62
Portafolio 2	-1.33	1.99	4.74
Portafolio 3	-1.22	2.23	4.94
Portafolio 4	-0.52	2.42	5.06