Pruebas de Contraste

Clase 8

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. Pruebas de Contraste
   - Inferencia sobre combinaciones lineales de medias
   - Pruebas de Hipótesis
   - Intervalos de Confianza
   - Combinaciones lineales de medias con varios factores
3. Ejemplo
   - Enunciado
   - Solución

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.

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	$\bar{Y}_{11..}$ $S_{11.}$	... ...	$\bar{Y}_{i1..}$ $S_{i1.}$	... ...	$\bar{Y}_{a1..}$ $S_{a1.}$	$\bar{Y}_{.1..}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	Nivel j	$\bar{Y}_{1j..}$ $S_{1j.}$	... ...	$\bar{Y}_{ij..}$ $S_{ij.}$	... ...	$\bar{Y}_{aj..}$ $S_{aj.}$	$\bar{Y}_{.j..}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	Nivel b	$\bar{Y}_{1b..}$ $S_{1b.}$	... ...	$\bar{Y}_{ib..}$ $S_{ib.}$	... ...	$\bar{Y}_{ab..}$ $S_{ab.}$	$\bar{Y}_{.b..}$
	Prom. columna	$\bar{Y}_{1...}$	...	$\bar{Y}_{i...}$	...	$\bar{Y}_{a...}$	$\bar{Y}_{....}$

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	$N-1$		

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.

Outline

Motivación

Hasta el momento sabemos cómo encontrar los factores que influyen sobre la variable de respuesta: Prueba ANOVA.

Una vez se han identificado que ciertas características influyen, entonces, si el problema lo requiere, se debe pensar en cómo seleccionar el nivel, o los tratamientos más convenientes.

Ejemplo: En el caso en que se quiere estudiar la influencia de la posición vetical del estante (arriba, centro o abajo) sobre las ventas de pan, la decisión final que se busca es encontrar el nivel en el cuál se vende más. Es decir, Optimizar la variable Y

Enfoques para Selección de Tratamientos

En general veremos dos procedimientos:

1. Inferencia sobre combinaciones lineales de medias.
2. Comparaciones múltiples de medias por pares.

Outline

Inferencia sobre combinaciones lineales de medias

En muchos experimentos, las decisiones se concentran en un sólo criterio que queda expresado como la combinación lineal de medias de tratamientos o de niveles de factores.

Ejemplo

Suponga que en el experimento para encontrar de qué depende la calidad de los tornillos que se producen, el tipo de aleación del acero (tipo 1 ó tipo 2) resulta ser un factor significativo. La decisión de trabajar con un tipo de acero, no necesariamente se da a partir de cuál de ellos maximiza el nivel de calidad, sino cual representa un mayor beneficio-costo.
Esto es, si el precio de venta del tornillo depende de la calidad, entonces, la decisión se toma con:

$$\theta = (p_1\mu_1 - b_1) - (p_2\mu_2 - b_2)$$

Si $\theta > 0$, entonces es más rentable trabajar con la aleación 1.

Inferencia sobre combinaciones lineales de medias

El problema de interés es hacer inferencia estadística (pruebas de hipótesis o intervalos de confianza) para combinaciones lineales de medias de tratamientos o de niveles de los factores, dado que sus valores poblacionales son desconocidos.

En el caso de un experimento con un sólo factor, el parámetro de interés puede ser escrito como:

$$\theta = \sum_{i=1}^{a} c_i \mu_i + K$$

donde $c_1, \cdots, c_a$ y K son las constantes conocidas dadas por el problema. A una expresión de este estilo se le denomina contraste.

Inferencia sobre combinaciones lineales de medias

El estimador natural de la media por nivel es su repetitivo promedio muestral por nivel (i.e $\hat{\mu}_i \to \bar{Y}_{i.}$), luego el estimador natural del contraste es la misma combinación lineal, pero con los promedios muestrales:

$$\hat{\theta} = \sum_{i=1}^{a} c_i \bar{Y}_{i.} + K$$

Desde la clase 1 estamos asumiendo que $Y_{ij} \sim N(\mu_i, \sigma^2)$ por lo tanto:

Estimador de Contraste

$$\hat{\theta} \sim Normal\left(\theta\ ,\ \sigma^2 \sum_{i=1}^{a} \frac{c_i^2}{n_i}\right)$$

donde n_i es el numero de datos por nivel.

Inferencia sobre combinaciones lineales de medias

Mejor estimador de la Varianza

Para hacer intervalos de confianza o pruebas de hipótesis, y dado que la varianza σ^2 es desconocida, se puede usar el mejor estimador que tenemos:

$$\hat{\sigma}^2 = MSE$$

y sus grados de libertad correspondientes para formar la distribución t.

Outline

Pruebas de Hipótesis

La hipótesis nula:

$$H_0 : \theta = \theta_0$$

Estadístico de prueba:

$$EP = \frac{\hat{\theta} - \theta_0}{\sqrt{MSE \left(\sum_{i=1}^{a} \frac{c_i^2}{n_i} \right)}} \sim t_{(gl_E)}$$

Región de rechazo:

H_1	RR
$H_1 : \theta \geq \theta_0$	$EP \geq t_{[1-\alpha, gl_E]}$
$H_1 : \theta \leq \theta_0$	$EP \leq t_{[\alpha, gl_E]}$
$H_1 : \theta \neq \theta_0$	$\lvert EP \rvert > t_{[1-\frac{\alpha}{2}, gl_E]}$

Outline

Intervalos de Confianza

Expresión General

$$IC\left(\theta, 1-\alpha\right) = \hat{\theta} \pm t_{[1-\frac{\alpha}{2};gl_E]}\sqrt{MSE\left(\sum_{i=1}^{a}\frac{c_i^2}{n_i}\right)}$$

En nuestro caso el intervalo toma la forma:

$$IC\left(\theta, 1-\alpha\right) = \left(\sum_{i=1}^{a} c_i\bar{Y}_{i.} + K\right) \pm t_{[1-\frac{\alpha}{2};gl_E]}\sqrt{MSE\left(\sum_{i=1}^{a}\frac{c_i^2}{n_i}\right)}$$

Outline

Combinaciones lineales de medias con varios factores

Cuando se tienen experimentos multifactoriales el procedimiento funciona muy similar, pero se debe tener en cuenta que se pueden definir combinaciones por celda, o por niveles de factores, o por grupos de celdas.

Por ejemplo, en el caso de dos factores, en donde se busca hacer inferencia sobre la combinación lineal de las medias por celda, entonces:

$$\theta = \sum_{i=1}^{a} \sum_{j=1}^{b} c_{ij} \mu_{ij} + K$$

Donde,

$$\hat{\theta} = \sum_{i=1}^{a} \sum_{j=1}^{b} c_{ij} \bar{Y}_{ij.} + K \sim Normal\left(\theta, \sigma^2 \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{c_{ij}^2}{n_{ij}}\right)$$

Dado que se asumen diseños balanceados ($n_{ij} = n$), siendo el número de réplicas por tratamiento.

Inferencia sobre combinaciones lineales de medias

Estadístico de Prueba

$$\frac{\hat{\theta}-\theta}{\sqrt{MSE\left(\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{c_{ij}^2}{n}\right)}} \sim t_{(gl_E)}$$

Intervalo de Confianza

$$IC\left(\theta, 1-\alpha\right) = \left(\sum_{i=1}^{a}\sum_{j=1}^{b} c_{ij}\,\bar{Y}_{ij.} + K\right) \pm t_{[1-\frac{\alpha}{2};gl_E]}\sqrt{MSE\left(\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{c_{ij}^2}{n}\right)}$$

Inferencia sobre combinaciones lineales de medias

En el caso de dos factores, si el parámetro de interés es la combinación de medias de un factor por nivel, entonces

$$\theta = \sum_{i=1}^{a} c_i \mu_{i.} + K$$

Donde,

$$\hat{\theta} = \sum_{i=1}^{a} c_i \bar{Y}_{i..} + K$$

Y se tiene que:

$$\frac{\hat{\theta} - \theta}{\sqrt{MSE\left(\sum_{i=1}^{a} \frac{c_i^2}{nb}\right)}} \sim t_{(gl_E)}$$

dado que se asumen diseños balanceados ($n_{ij} = n$).

Outline

Ejemplo

Se busca determinar los factores que afectan la calidad de un tornillo. Para eso se toman como Factor A: Aleación (a=2), Factor B: Velocidad troquelado (b=3), Factor C: Temperatura troquelado (c=2). Para cada tratamiento se toman 5 muestras. Al desarrollar el diseño experimental y encontrar el mejor modelo se obtiene la siguiente tabla ANOVA

Fuente	SS	gl	MS	F	Pvalor
A	540360.60	1	540360.60	695.20	0.00
B	49319.63	2	24659.82	31.73	0.00
C	382401.67	1	382401.67	491.98	0.00
Error	42750.03	55	777.27		
Total	1014831.93	59			

Ejemplo

Adicionalmente se tiene la siguiente información sobre el factor aleación:

	Media	**Desviación**	**IC -**	**IC +**
Aleación 1	1155.93	16.84	1122.23	1189.63
Aleación 2	966.133	16.17	933.76	998.5

Se sabe que la aleación 1 tiene un costo de material por unidad de \$1000, mientras que para la aleación 2 es de \$700. Se sabe que el precio que se paga por tornillo depende de la calidad (Y) como: $Precio = 2 + Y$. Realice una prueba que le permita determinar con cuál de las dos aleaciones se deben fabricar los tornillos. Realice también el intervalo de confianza correspondiente.

Outline

Solución

Del enunciado sabemos qué:

- Se va a tomar la decisión basado únicamente en el primer factor (Tipo de Aleación). Es decirt voy a tener que utilizar los promedios muestrales referentes a dicho factor (i.e $\mu_{i\ldots}$)
- Sabemos que de forma general la Utilidad se define como $Ingresos - Costos$. En este caso la puedo expresar como:

$$Utilidad_i = (2 + \mu_{i\ldots}) - Costo_i \quad \forall i \in \{1, 2\}$$

Así para cada aleación la utilidad queda como:

$$\text{Aleación } 1 : (2 + \mu_{1\ldots}) - 1000 \quad \text{y} \quad \text{Aleación } 2 : (2 + \mu_{2\ldots}) - 700$$

Solución

Como queremos comparar cual de las dos utilidades resulta mayor, podemos escribir nuestro contraste como:

$$\theta = (2 + \mu_{1...} - 1000) - (2 + \mu_{2...} - 700)$$

$$\theta = (\mu_{1...} - 998) - (\mu_{2...} - 698)$$

El estimador del contraste sería, la misma combinación lineal pero reemplazando las medias poblacionales con su respectivo promedio muestral, es decir:

$$\hat{\theta} = \left(\bar{Y}_{1...} - 998\right) - \left(\bar{Y}_{2...} - 698\right)$$

Como asumimos que $Y_{ijkl} \sim N(\mu_{ijkl}, \sigma^2)$ Entonces:

$$\hat{\theta} \sim N\left(\theta, Var(\hat{\theta})\right)$$

Solución

La varianza del estimador resulta:

$$Var(\hat{\theta}) = Var\left(\left(\bar{Y}_{1\ldots} - 998\right) - \left(\bar{Y}_{2\ldots} - 698\right)\right)$$

$$Var(\hat{\theta}) = Var\left(\bar{Y}_{1\ldots} - 998\right) + Var\left(\bar{Y}_{2\ldots} - 698\right) - 2Cov(\bar{Y}_{1\ldots} - 998, \bar{Y}_{2\ldots} - 698)$$

El término de la covarianza es cero ya que las medias de niveles/tratamietos diferentes son independientes una de la otra (Supuesto del ANOVA), entonces:

$$Var(\hat{\theta}) = Var\left(\bar{Y}_{1\ldots} - 998\right) + Var\left(\bar{Y}_{2\ldots} - 698\right)$$

$$Var(\hat{\theta}) = Var\left(\bar{Y}_{1\ldots}\right) + Var\left(998\right) + Var\left(\bar{Y}_{2\ldots}\right) + Var\left(698\right)$$

$$Var(\hat{\theta}) = Var\left(\bar{Y}_{1\ldots}\right) + Var\left(\bar{Y}_{2\ldots}\right)$$

De proba 1 sabemos qué sí $X \sim N(\mu, \sigma^2)$ Entonces $\bar{X} \sim N(\mu, \frac{\sigma^2}{n^*})$ donde n^* es el número de datos bajo el cual se hace el promedio

Solución

Entonces:

$$Var(\hat{\theta}) = \frac{\sigma^2}{n^*} + \frac{\sigma^2}{n^*}$$

En este caso como tengo un experimento de 3 factores y estoy haciendo inferencia sobre el factor 1, el número de datos por cada nivel del factor aleación es bcn (b número de niveles del factor 2, c número de niveles del factor 3, n número de réplicas o número de datos por tratamiento).

$$Var(\hat{\theta}) = \frac{2\sigma^2}{bcn}$$

Así la distribución de $\hat{\theta}$ es:

$$\hat{\theta} \sim N\left(\theta, \frac{2\sigma^2}{bcn}\right)$$

Solución

Ya qué no conocemos la varianza σ^2 , debemos estimarla. Para esto y ya qué estamos en el contexto de un diseño experimental utilizamos el *MSE* pues sabemos que es nuestro mejor estimador.

Las desviaciones muestrales por nivel, que nos da la tabla 2 NO sirven como estimador de la varianza pues estas no tienen en cuenta la información de diseño experimental realizado.

Combinando el estimador $\hat{\theta}$ con la estimación de la varianza ($\hat{\sigma}^2 = MSE$) obtenemos un estadístico:

$$\frac{\hat{\theta} - \theta_0}{\sqrt{MSE\left(\frac{2}{bcn}\right)}} \sim t_{(gl_E)}$$

En nuestro caso particular podemos expresar la pregunta de interés con las siguientes hipótesis: $H_0 : \theta = 0$ y $H_1 : \theta \neq 0$

Solución

Combinando todo la prueba queda:

HIPÓTESIS

$$H_0 : \theta = 0$$

$$H_1 : \theta \neq 0$$

ESTADÍSTICO DE PRUEBA

$$EP = \frac{\hat{\theta} - \theta_0}{\sqrt{MSE\left(\frac{2}{bcn}\right)}} \sim t_{(gl_E)}$$

$$EP = \frac{(1155.93 - 998) - (966.13 - 698) - 0}{\sqrt{777.27\left(\frac{2}{3*2*5}\right)}} \sim t_{(55)}$$

$$EP = -15.31$$

REGIÓN DE RECHAZO

$$t_{1-\alpha/2,gle} \rightarrow t_{0.975,55} = 2.004$$

Solución

CONCLUSIÓN

Como $|EP| > t_{1-\alpha/2,glE}$ se rechaza la hipótesis nula por ende existen diferencias en las utilidades recibidas por los tornillos de la aleación 1 y la aleación 2

¿Cuál es la aleación que se debe seleccionar para hacer los tornillos?