Múltiples Contrastes

Clase 9

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

Pruebas de Contraste: 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 influyes, 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) sober 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

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.

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

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

Motivación

A la hora de realizar pruebas posteriores al ANOVA, existe la posibilidad que para contestar UNA pregunta en particular, sea necesario realizar de más de un contraste.

Ejemplo

Hallar la posición vertical del estante de la tienda (arriba, centro o abajo) que optimice las ventas de un producto, siendo esta la variable Y de interés.

Aca, para hallar el óptimo, se deben comparar todos los tratamientos entre ellos y así identificar cual es el que genera mayor ventas. Es decir, se deben realizar varios contrastes.

¿Ven algo raro en esto? ¿Cómo enfrentarían este problema?

Motivación

Toda pregunta de este estilo se contesta por medio de una prueba de hipótesis, en donde se fija un nivel de significancia para controlar el error Tipo I.

Al hacer múltiples pruebas estadísticas para contestar una pregunta, el nivel de error se acumula, luego el nivel de significancia real asociado a nuestra pregunta no coincide exactamente con el que se hace individualmente.

Ejemplo

Suponga que quiere medir la distancia horizontal entre 2 locaciones de Bogotá. Para esto divide el trayecto en 10 segmentos disjuntos, y se mide su distancia. El error de medición en cada uno de los segmentos es de ± 1 metro. Cuanto sería el error para el total del trayecto? ± 10 metros.

Lo mismo sucede con la significancia total de la prueba.

Motivación

La significancia está asociada a equivocase al rechazar H_0 cuando realmente es verdad, al hacer una prueba estadística a un contraste en particular. Al hacer varias pruebas, el nivel de significancia total corresponde a equivocarse en alguna de las pruebas, luego la posibilidad de equivocase es más grande.

Matemáticamente, si se realizan m pruebas de hipótesis independientes, entonces:

$$\alpha_m = 1 - (1 - \alpha_I)^m$$

donde α_I es el nivel de significancia con el que se hace cada contraste, y α_m es el el nivel de significancia real de las m pruebas.

Si queremos fijar un nivel de significancia para contestar nuestra pregunta, no es adecuado usar el mismo nivel de significancia en cada una de las pruebas individuales. ¡Debemos hacer algo para que se mantenga el nivel de significancia que queremos para contestar la pregunta!

Outline

Comparaciones múltiples por pares

Comparaciones múltiples por pares

Cuando hacemos ANOVA queremos responder la pregunta:

$$H_0: \quad \text{El Factor NO influye sobre } Y \Leftrightarrow \mu_1 = \mu_2 = \cdots = \mu_a$$
$$H_1: \quad \text{El Factor SI influye sobre } Y \Leftrightarrow \text{Algún par } \mu_i \neq \mu_j$$

Y sí concluimos que el factor es significativo, entonces existe por lo menos un par de medias de los niveles del factor que son diferentes entre si. La pregunta ahora se vuelve, ¿Cuales medias son diferentes entre si? ¿Cuales medias son "óptimas"según el contexto?

Comparaciones múltiples por pares

La hipótesis correspondiente es:

$$H_0^{ij} : \mu_i = \mu_j$$

$$H_1^{ij} : \mu_i \neq \mu_j$$

$$\forall i, j \in \{1, ..., a\}, i \neq j$$

Observe que esto da un total de $m = \binom{a}{2}$ pruebas de hipótesis diferentes que se deben probar de forma simultanea para contestar la pregunta.

El objetivo es realizar las m pruebas de tal forma que la significancia total se mantenga en un valor de α que nosotros mantengamos fijo y controlado (i.e $\alpha_m = \alpha$).

Se han desarrollado diferentes enfoques para este propósito, nos enfocaremos en el procedimiento de Tukey.

Prueba de Tukey

El procedimiento propuesto por Tukey se basa en la distribución del rango studentizado:

$$q = \frac{\bar{Y}_{max} - \bar{Y}_{min}}{\sqrt{\frac{MSE}{2}(\frac{1}{n_{max}} + \frac{1}{n_{min}})}} \sim q_{m,glE}$$

donde $\bar{Y}_{max} = max\{\bar{Y}_{i.} - \mu_i\}$, $\bar{Y}_{min} = min\{\bar{Y}_{i.} - \mu_i\}$ y n_{max}, n_{min} son sus correspondientes tamaños de muestra. Esta distribución se encuentra tabulada. Si lo piensan, bajo H_0 se tiene que:

$$P(\bar{Y}_{max} - \bar{Y}_{min} \leq K) = P(|\bar{Y}_{i.} - \bar{Y}_{j.}| \leq K, \forall i, j \in \{1, ..., a\}, i \neq j)$$

Luego, podemos utilizar la distribución del rango studentizado para nuestras comparaciones por pares!

Prueba de Tukey

En ese orden de ideas, el procedimiento de Tukey sugiere decretar diferencias entre un par de medias de los niveles del factor con un nivel de significancia total de α si:

$$|\bar{Y}_{i.} - \bar{Y}_{j.}| \geq q_{\alpha,m,glE}\sqrt{\frac{MSE}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

donde $q_{\alpha,m,glE}$ es el valor crítico de la tabla (con m siendo el número de medias a comparar y glE los grados de libertad del error), y n_i es el numero de datos en el nivel i. En el caso en que el diseño es balanceado, el término a la derecha se simplifica y queda:

$$|\bar{Y}_{i.} - \bar{Y}_{j.}| \geq q_{\alpha,m,glE}\sqrt{\frac{MSE}{n}}$$

Observe que dicho término es constante para todas las comparaciones.

Prueba de Tukey

Otras formas equivalentes de presentar esta región de rechazo se dan por medio del estadístico q o de un intervalo de confianza:

$$q = \frac{|\bar{Y}_{i.} - \bar{Y}_{j.}|}{\sqrt{\frac{MSE}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \geq q_{\alpha,m,glE}$$

$$|\bar{Y}_{i.} - \bar{Y}_{j.}| \pm q_{\alpha,m,glE}\sqrt{\frac{MSE}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \text{ Si NO incluye el } 0$$

También se puede hacer para los tratamientos, cambiando el número de comparaciones que se debe hacer m, así como utilizando las medias por tratamiento $\bar{Y}_{ij.}$ y su respectivo tamaño de muestra n_{ij}

Outline

Optimización

Suponga que se quiere hallar el tratamiento donde obtiene el mínimo o el máximo para la variable de interés bajo un nivel de significancia dado.

Es claro que los candidatos son los tratamientos donde se tiene tanto el mínimo como el máximo promedio muestral respectivamente.

Como una consecuencia del proceso de Tukey que se acaba de realizar, se tiene entonces que aquellos tratamientos que sean estadísticamente iguales a los candidatos, son también tratamientos que alcanzan el mínimo y máximo de la variable de interés.

Para tener en cuenta

En el caso que hayan interacciones significativas en el modelo, es necesario realizar el procedimiento de Tukey por tratamiento. Esto debido a que el efecto de la interacción no se puede ver individualmente.

Outline

Enunciado

Se hizo un estudio para comparar el rendimiento de tres marcas de gasolina competidoras. Se seleccionaron al azar cuatro modelos de automóvil de tamaño variable. A continuación se presentan los datos, en millas por galón.

Gasolina		
A	**B**	**C**
32.4	32.6	38.7
28.8	28.6	39.9
29.5	37.6	39.1
34.4	34.2	37.9

a. ¿Es el tipo de gasolina un factor significativo?

b. Determine cuales pares de medias son diferentes entre sí.

Outline

Solución

Para responder el literal a. debemos hacer un ANOVA de un factor. Utilizando los datos del enunciado sabemos que: $a = 3$, $n = 4$ y $N = 12$. Con esto, se calculan las respectivas sumas de cuadrados y da como resultado:

Fuente	**SS**	**gl**	**MS**	**F**	**pvalor**
Gasolina	125.29	2	62.64	8.774	0,008
Error	64.26	9	7.14		
Total	189.54	11			

Del diseño de un factor podemos concluir que existe al menos un par de medias que son diferentes entre sí. La pregunta ahora se vuelve ¿Cuáles son los pares diferentes?

Solución

Utilizando el procediminto de Tukey tendríamos 3 pares de hipotesis:

$$H_0 : \mu_a = \mu_b, \quad H_0 : \mu_a = \mu_c, \quad H_0 : \mu_b = \mu_c$$

$$H_1 : \mu_a \neq \mu_b \quad , H_1 : \mu_a \neq \mu_c, \quad H_1 : \mu_b \neq \mu_c$$

Cuyos estimadores serían $\bar{Y}_{A.} = 31.27$, $\bar{Y}_{B.} = 33.25, \bar{Y}_{C.} = 38.9$.

Para determinar diferencias entre cada par de medias debemos calcular el valor absoluto de la diferencia ($|\bar{Y}_{i.} - \bar{Y}_{j.}|$) y compararlo contra el punto critico dado por $q_{\alpha,m,glE} * \sqrt{\frac{MSE}{n}}$. Calculando para cada par de medias se obtiene:

$$|\bar{Y}_{A.} - \bar{Y}_{B.}| = 1.975 \quad |\bar{Y}_{A.} - \bar{Y}_{C.}| = 7.625 \quad |\bar{Y}_{B.} - \bar{Y}_{C.}| = 5.65$$

y un punto crítico de: $q_{0.05,3,9} * \sqrt{\frac{7.14}{4}} = 5.274$

Solución

Haciendo la comparación respectiva se concluye qué:

- Las medias de la gasolina Tipo A y Tipo B son iguales (No rechazo H_0 pues $1.975 \leq 5.274$).
- Las medias de la gasolina Tipo A y Tipo C son diferentes (Rechazo H_0 pues $7.625 \geq 5.274$).
- Las medias de la gasolina Tipo B y Tipo C son diferentes (Rechazo H_0 pues $5.65 \geq 5.274$).

Si se quiere maximizar el rendimiento ¿Qué tipo de gasolina se debe escoger? y ¿Si se quiere minimizar?