IN 540-01
Métodos Estadísticos para Economía y Gestión
2005/2

SOLUCIÓN CONTROL N° 1
22/09/2005

1. El salario medio (expresado en miles de $) de los trabajadores de un sector es de 180 y la desviación típica es 40. Además, el 15% de los trabajadores gana más de 200. Si se consideran muestras aleatorias de 100 trabajadores:

   a) ¿En qué porcentaje de muestras saldrá un sueldo medio menor que 170? *(0.5 ptos).*
   b) ¿En qué porcentaje de muestras saldrá más del 20% de trabajadores con sueldo superior a 200? *(0.5 ptos).*
   c) Repita a) y b) anteriores para n = 400 y compara los resultados *(0.5 ptos).*
   d) ¿Por qué $\bar{x}$ (salario medio) y p (proporción de trabajadores que gana más de 200) tienen el carácter de variables aleatorias? Indica otras posibles variables aleatorias muestrales *(0.5 ptos).*

   Solución

   a) Si aplicamos la aproximación que proporciona el TCL, entonces dado $\bar{X}$ en salario medio muestral, entonces

$$P\left[\left(\frac{\bar{X}-180}{40/\sqrt{100}} < \frac{170-180}{4}\right)\right] = P[Z<-2.5]$$

   donde $Z \sim N(0,1)$. De la tabla normal proporcionada $P[Z<-2.5]=0.5-0.494=0.006$, lo cual indica que en 0.6% de las muestras aleatorias de 100 trabajadores, se observará un salario medio menor que 170.

   b) Si aplicando la misma aproximación sobre la proporción muestral, se tiene

$$P\left[\left(\frac{p-0.15}{\sqrt{.015\times 0.85}/\sqrt{100}} > \frac{0.2-0.15}{0.1129}\right)\right] = P[Z>0.4428]$$

   donde $Z \sim N(0,1)$. De la tabla normal proporcionada $P[Z>0.4428]=1-(0.5+0.170)=0.330$, lo cual indica que en 33% de las muestras aleatorias de 100 trabajadores, saldrá más de 20% de trabajadores con sueldo superior a. 200.

   c) Se aplican las misma fórmulas, excepto que aparece $\sqrt{400}=20$ en vez de $\sqrt{100}=10$ en los errores estándar del salario medio muestral y la proporción de trabajadores en la muestra con sueldo superior a 200. En el primer caso, se debe calcular ahora

$$P\left[Z<-\frac{2.5}{2}\right]=0.5-0.394=0.106$$

   o sea un 10,6% de las muestras de 400 casos mostrará salarios medios menores que 170.
   En el segundo caso,

$$P\left[Z>\frac{0.4428}{2}\right]=1-(0.5+0.087)=0.413$$

   lo cual indica que en 41.3% de las muestras aleatorias de 400 trabajadores, saldrá más de 20% de trabajadores con sueldo superior a. 200.

   d) Las variables anteriores son aleatorias debido a que la muestra, según el enunciado, lo es. Cualquier otra función de la muestra de 100 observaciones, un estadístico, será también una variable aleatoria: los estadísticos de orden, la mediana, etc.

2. Se han publicado resultados de un estudio sobre precios de combustibles en Santiago. En distintas zonas de la capital (Norte, Oriente, Poniente y Sur), se escogió aleatoriamente servicentros de seis distribuidoras, observándose los siguientes precios (en $) del petróleo diesel:

	Zona			
Distribuidora	Norte	Sur	Oriente	Poniente
Shall	228	242	586	272
Essa	352	232	560	328
Copel	448	446	566	234
YQS	366	476	484	306
Tejano	318	236	536	304
Plano Blanco	298	444	368	216

Se le pide probar al nivel del 5% la hipótesis que en promedio, en cada zona de Santiago se paga el mismo precio por el petróleo diesel.

a) ¿Qué necesita para evaluar la aplicabilidad de ANOVA al problema? ¿cómo haría esa evaluación estadísticamente? (*0.5 ptos).*
b) Use ANOVA de una vía para resolver el problema *(1.0 ptos).*
c) Ahora, determine si la política de precios de cada distribuidora afecta o no los precios promedio observados en cada zona *(0.5 ptos)*.

Solución

a) Se necesita verificar dos supuestos básicos: que los precios en cada zona siguen una distribución normal y que se tiene homocedasticidad (las varianzas de los precios según zona son similares). Para confirmar estos supuestos se puede aplicar algún test de normalidad (por ej. Uno basado en Chi-cuadrado) y algún otro de homocedasticidad, como el test de Levene.

b) Calculamos las medias básicas:

	Zona				
Distribuidora	Norte	Sur	Oriente	Poniente	$\overline{X}_i$
Shall	228	242	586	272	332.0
Essa	352	232	560	328	368.0
Copel	448	446	566	234	423.5
YQS	366	476	484	306	408.0
Tejano	318	236	536	304	348.5
Plano Blanco	298	444	368	216	331.5
$\overline{X}_j$	335.0	346.0	516.7	276.7	368.6
					$\overline{\overline{X}}$

En seguida, se debe calcular las sumas de cuadros de interés:

$$SCT = \sum_{i=1}^{6}\sum_{j=1}^{4}(X_{ij} - \overline{\overline{X}})^2 = \ 334{,}059.8$$

$$SCTR = \sum_{j=1}^{6} 6(\overline{X}_j - \overline{\overline{X}})^2 = \ 192{,}091.2$$

$$SCE = SCT - SCTR = \ 141{,}968.7$$

Luego, los cuadrados medios y F

$$CMTR = \frac{SCTR}{23} = 64{,}030.4 \quad , \quad CME = \frac{SCE}{20} = 7{,}098.4 \quad \text{y} \quad F = \frac{CMTR}{CME} = 9.0$$

En tabla de valores críticos al 5% de la distribución de F proporcionada, no alcanza a aparecer $n_1=23$ ni tampoco $n_2=20$. Pero los valores extremos de la tabla indican que el valor crítico es menor que 3.0, con lo cual el valor F observado, cuyo valor es 9.0, será significativamente distinto de 0 al 5%. Con ello, se rechaza la hipótesis que se paga lo mismo por el petróleo en las distintas zonas de Santiago.

c) Lo que se pide es, en pocas palabras, bloquear por los distintos tipos de distribuidora. Así, interesa las Cuma de cuadrados por bloque

$$SCBL = \sum_{i=1}^{4} 4(\overline{X}_i - \overline{\overline{X}})^2 = 30,746.8$$

la nueva suma de cuadrados residuales

$$SCE = SCT - SCTR - SCB = 111,221.8$$

y los nuevos cuadrados medios y el primero de los nuevos F (es decir, aquel asociado al bloqueo)

$$CME = \frac{SCE}{5 \times 3} = 7,414.8 \quad , \; CMBL = \frac{SCBL}{5} = 6,149.4 \quad \text{y} \; F = \frac{CMBL}{CME} = 0.83$$

En tabla de valores críticos al 5% de la distribución de F proporcionada, con el mismo criterio anterior, se tiene que el F observado es demasiado bajo para rechazar la hipótesis que política de precios de cada distribuidora no afecta los precios promedio observados. Así, se debe considerar que dicha política no afecta los precios promedio observados en cada zona.

3. Responda breve pero justificadamente (en no más de 10 líneas c/u):

a) Cuando la varianza inter-estratos es grande, ¿el muestreo aleatorio estratificado se desempeña mejor o peor que el muestreo aleatorio simple? *(0.5 ptos)*

b) ¿Los Análisis Discriminante y de Clusters son complementarios o suplementarios? *(0.5 ptos)*

c) En el ejemplo de Análisis Factorial Discriminante visto en clases, la calidad del vino del valle del Maipo en función de variables climatológicas, arrojó las siguientes funciones discriminantes, estadísticos de interés y valores estimados:

$$Z_1 = 0.9984 \times Temperatura^* + 0.8329 \times Sol^* - 0.5138 \times Lluvia^*$$

$$Z_2 = -0.0017 \times Temperatura^* - 0.5965 \times Sol^* + 0.5831 \times Lluvia^*$$

	Media	Desv. Típica
Temperatura	3157.9	141.2
Sol	1247.3	126.6
Lluvia	360.4	91.4

Orden	η_m^2	Λ_{3-m}	*F*	*p-value*
1	0.763	0.219	10.98	0.0001
2	0.075	0.925	1.22	0.3089

Funciones Discriminantes (Análisis Factorial Discriminante)			
Año	Calidad	Z_1	Z_2
64	2	-0.88	-0.87
65	3	-2.33	-0.09
66	2	-0.99	0.83
67	3	-2.73	0.25
68	1	0.74	1.70
69	1	2.23	0.48
70	3	-2.75	1.11
71	3	-2.53	0.24
72	3	-3.73	2.11
73	2	1.13	1.37
74	1	2.17	-0.04
75	3	-0.36	1.36
76	3	-2.02	-0.54
77	1	1.55	-0.53
78	2	-0.73	-0.79
79	2	-0.31	-1.8
80	2	0.34	-1.56
81	3	-2.45	0.8
82	2	0.79	0.16
83	1	2.41	-0.46
84	2	1.14	-0.82
85	1	3.54	-0.93
86	2	-0.55	-1.1
87	1	3.18	1.95
88	2	0.21	-1.28
89	1	4.12	1.22
90	2	1.47	0.22
91	3	-1.68	0.23
92	1	2.17	-0.49
93	1	0.35	0.13
94	3	-2.10	-1.49
95	1	0.87	-0.05
96	3	-1.09	-0.98
97	3	-1.18	-0.34

i. ¿Son las dos funciones discriminantes realmente utilizables? *(0.5 ptos)*

ii. Con una regla de decisión adecuada, ¿Cómo diría usted que viene el vino del año 1998, si Temperatura = 3000, Sol = 1100, Lluvia = 300. *(0.5 ptos)*

Solución

a) Consideremos el caso de muestreo estratificado proporcional. Si la varianza interclase es pequeña, entonces la varianza total se parece a la varianza intraclase, con lo cual ambos diseños entregan errores similares. En cambio si la varianza interclase es grande, se tendrá mejores resultados en cuanto a errores menores para un mismo tamaño muestral en el caso de muestreo estratificado que en el muestreo aleatorio simple, por cuanto esto significa que la variable de estratificación está asociada (quizás fuertemente) a la variable de interés, lo cuál contribuye a una estimación más precisa al interior de cada estrato (la varianza intraclase es relativamente pequeña si la varianza interclase es grande), y por ende en el total.

b) Pueden considerarse complementarios, pero no sustitutos o suplementarios, ya que en el Análisis Discriminante (AD) los grupos (clusters) están determinados a priori y el objetivo es obtener la combinación lineal de variables independientes que discrimina mejor entre los grupos. En Análisis de Cluster (AC), los grupos (clusters) no están predeterminados y, de hecho, el objetivo es determinar la mejor manera en que los "casos" se clasifican en grupos. Un AD puede usarse para validar un AC.

c) Se tiene que:

i. El cuadro con los Λ de Wilks indica que sólo la función discriminante Z_1 tiene un significativo poder discriminante (p-value es prácticamente nulo). No ocurre los mismo con Z_2, que posee un escaso poder discriminante (Λ de Wilks alto y F pequeño). Luego, sólo Z_1 es realmente utilizable.

ii. Los * en las variables que independientes que definen Z_1 y Z_2 indican que éstas están estandarizadas. Usando i., se tiene que la regla de clasificación de nuevo caso estará dada, en este caso, por el valor que alcance:

$$Z_1 = 0.9984 \times \left(\frac{3000 - 3157.9}{141.2} \right) + 0.8329 \times \left(\frac{1100 - 1247.3}{126.6} \right) - 0.5138 \times \left(\frac{300 - 360.4}{91.4} \right) = -1.74603$$

Observando en el cuadro de Funciones Discriminantes proporcionado, se observa que $Z_1 < -1$, se tiene en la gran mayoría de casos en que la Calidad del Vino es categoría 3. Luego, se predice que el nuevo caso se clasifica con Calidad del Vino = 3.