CURSO: IN540-02 MÉTODOS ESTADÍSTICOS PARA ECONOMÍA Y GESTIÓN
PROFESORES: MATTIA MAKOVEC - MARCELO HENRÍQUEZ
P. AUXILIAR: DARÍO ZUÑIGA
SEMESTRE: OTOÑO 2008

Corrección Control Nº 1

Observaciones:

1. *La mayoría de las preguntas requieren una respuesta directa, donde lo importante es la justificación. Si se entrega un resultado sin justificarlo adecuadamente, no se asigna puntaje a la respuesta. En negrita se anotará el argumento central para cada respuesta.*
2. *En las preguntas que requieren desarrollo estadístico, se destacará del mismo modo los aspectos clave que llevan a un resultado correcto.*

P1. Debe justificar brevemente sus respuestas.

a) (10 ptos.)¿Cuál es el error muestral en la estimación de una proporción poblacional p, en el caso de α=5%, n=1250, N finito pero muy grande y varianza máxima?

Solución:

Para responder es necesario saber de qué tipo de muestreo se trata. Los más directo es suponer que se trata de un MAS. *La expresión a partir de la cual hay que encontrar la fórmula del error muestral, considerando que n=1250 es suficientemente grande para utilizar la aproximación normal, es* $P\left(\left|\frac{(\bar{x}-\mu)}{\sigma_{\bar{x}}}\right| \le \varepsilon\right) = 1-\alpha$, *donde* ε, *corresponde al error muestra. De aquí:*

- $\varepsilon = z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)}$, *si* σ^2, *la varianza poblacional de x, es conocida* *(1.1)*
- $\varepsilon = z_{\alpha/2}\sqrt{\frac{s^2}{n}\left(\frac{N-n}{N}\right)}$, ***si*** σ^2 ***, la varianza poblacional de x, es desconocida y se estima por*** $s^2\left(\frac{N-1}{N}\right)$ ***(1.2)***

Para estimar una proporción poblacional p se considera $\hat{s}^2 = \hat{p}(1-\hat{p})$ ***y como se habla de varianza máxima, esto significa que se puede utilizar*** $\hat{p} = 0.5$***. Además N es muy grande, lo cual indica que*** $\frac{N-n}{N} \approx 1$***. Por último, hay que recordar que*** $z_{\alpha/2}$ ***=1.96 cuando α=5%. Así, utilizando (1.2), se tiene*** $\varepsilon = 1.96\sqrt{\frac{0.25}{1250}} == 0.0277$ ***, es decir 2.8%.***

b) (5 ptos.) Cuando la varianza inter-estratos es grande, ¿el muestreo aleatorio estratificado se desempeña mejor o peor que el muestreo aleatorio simple?

Solución:

Consideremos el caso de muestreo estratificado proporcional. ***Si la varianza interestratos (o 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.***

c) (5 ptos.) ¿Cuál diría usted que aspecto metodológico común y más relevante en el diseño de los análisis de Clusters, Discriminante y ANOVA?

Solución:

Vimos que el enfoque que atraviesa el diseño de los métodos mencionados es una adecuada (y específica al método) descomposición de la variación de las datos en análisis. En el caso de Clusters, Discriminante y ANOVA, donde

tenemos una(s) variable(s) nominal(es) o de categorías, se utilizan las variaciones Intra-grupo e Inter-grupo. Entonces, el aspecto relevante dice relación con una descomposición adecuada de la variación de los datos.

d) (5 ptos) ¿Por qué generalmente el método K-medias produce diferentes soluciones, dependiendo de la secuencia de las observaciones en la base de datos? ¿Qué recomienda para corregir este problema?

Solución:

Esto ocurre porque para partir, K-medias requiere de centros iniciales de los clusters. Si estos centros estás tomados de los datos (por ejemplo los primeros k datos, si se buscan k grupos, que es el defecto en SPSS), entonces los grupos obtenidos estarán influidos por esos casos, en el sentido que si esos datos iniciales son otros, podrían obtenerse grupos diferentes. Para resolver la dificultad, se puede aleatorizar los casos (mezclarlos aletariamente) y aplicar el método sobre distintos ordenamientos de los datos.

e) (5 ptos.) Sea $D_i^2 = (x_i - \bar{x})^t S^{-1} (x_i - \bar{x})$ la distancia de Mahalanobis de una observación *i* al vector de media muestral. Muestre que $\sum_{i=1}^{n} D_i^2 = np$ siendo *p* la dimensión de los vectores de datos y *n* el tamaño de la muestra.

Solución:

El elemento clave es notar que $D_i^2 = (x_i - \bar{x})^t S^{-1} (x_i - \bar{x}) = traza[S^{-1}(x_i - \bar{x})(x_i - \bar{x})^t]$.

Por lo tanto, $\sum_{i=1}^{n} D_i^2 = traza[S^{-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^t]$ ***y como*** $\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^t = nS$ **, entonces**

$\sum_{i=1}^{n} D_i^2 = traza[nS^{-1}S] = traza[nI_p] = np$ **.**

P2. Las negociaciones entre los trabajadores y la gerencia de una empresa minera se concentran en incentivos remuneracionales asociados a la productividad del trabajador. A los trabajadores se les puede pagar con participación en las utilidades, comisiones, salarios o un plan de bonificaciones. Tres profesionales y tres operarios, seleccionados al azar, se trataron con métodos distintos de pago. Su productividad diaria, en una medida de unidades convencionales, aparece en el cuadro siguiente:

	Profesionales			Operarios		
Salarios	4	5	3	4	3	4
Comisiones	8	7	7	6	7	5
Participaciones	8	9	8	7	8	6
Bonificaciones	6	5	5	4	5	3

a) (30 ptos) Utilice ANOVA de dos vías (con un nivel del 1%) para analizar los efectos relativos de las formas de pago involucradas.

Solución

Para una adecuada solución, conviene partir del modelo que se plantea para el problema:

$$y_{ijk} = \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \quad (1)$$

Consideremos que α_i ***captura el efecto central que se quiere medir, correspondiente a las formas de pago;*** β_j ***el efecto del tipo de trabajador (profesional u operario) y*** γ_{ij} ***la interacción de ambos. Así, en la matriz de datos se tiene: i= 1,…,I con I=4; j=1,…,J con J=2 y k=1,..,K con K=3.***

Comparando con en el ANOVA básico visto en clases, podemos suponer que en este caso se está considerando y_{ijk} *como la variable de resultados centrada, es decir si* p_{ijk} *es la productividad del trabajador ijk, entonces* $y_{ijk} = p_{ijk} - \mu$*, obteniendo un modelo equivalente a (1):*

$$p_{ijk} - \mu = \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

***** Ahora bien, si*** μ_i ***es la media poblacional de productividad de la forma de pago i;*** ν_j ***es la media poblacional de la productividad del tipo de trabajador j; y*** η_{ij} ***la media poblacional del tipo de trabajador j cuando se le paga en la forma i, entonces:***

$$p_{ijk} - \mu = (\mu_i - \mu) + (\nu_j - \mu) + (\eta_{ij} - (\mu_i + \nu_j - \mu)) + (p_{ijk} - \eta_{ij}) \quad \textbf{(2)}$$

de donde se tiene que $\alpha_i = (\mu_i - \mu)$, $\beta_j = (\nu_j - \mu)$ **y** $\gamma_{ij} = (\eta_{ij} - (\mu_i + \nu_j - \mu))$**. Las medias poblacionales se pueden estimar por las medias maestrales siguientes:**

μ **:** $\quad \overline{p} = \frac{1}{n}\sum_i \sum_j \sum_k p_{ijk}$ **con *n = IJK = 24.***

μ_i : $\quad \overline{p}_{i+} = \frac{1}{JK}\sum_j \sum_k p_{ijk}$

ν_j : $\quad \overline{p}_{+j} = \frac{1}{IK}\sum_i \sum_k p_{ijk}$

η_{ij} : $\quad \overline{p}_{ij} = \frac{1}{K}\sum_k p_{ijk}$

La estimación del modelo sería $p_{ijk} - \overline{p} = (\overline{p}_{i+} - \overline{p}) + (\overline{p}_{+j} - \overline{p}) + (\overline{p}_{ij} - (\overline{p}_{i+} + \overline{p}_{+j} - \overline{p})) + (p_{ijk} - \overline{p}_{ij})$ ***(3)***

Se observa que (3) es de la forma T = A + B + C + E. Con un poco de álgebra se llega a la descomposición de las sumas de cuadrados:

$$SC\_T = SC\_A + SC\_B + SC\_C + SC\_E$$

****

*(Desarrollo del modelo ** : 10 puntos)*

**** Con estas sumas de cuadrados se obtiene la tabla de ANOVA siguiente* ***(¡¡hay que hacer cálculos!!):***

	SC	***g.l.***	***MC***	***F***
FACTOR (A)	*56,13*	*I – 1 = 3*	*18,71*	*28,06*
FACTOR (B)	*7,04*	*J – 1 = 1*	*7,04*	*10,56*
INTERACCIÓN (C)	*1,13*	*(I – 1)(J – 1) = 3*	*0,38*	*0,56*
ERROR (E)	*10,67*	*n – IJ = 16*	*0,67*	
TOTAL (T)	*74,96*	*n – 1 = 23*		

*Puede hallarse variación en las respuestas a nivel de décimas o centésimas, según el número de cifras significativas que se use en los cálculos. El tema de los grados de libertad es sencillo en un diseño balanceado.****
*(Cuadro ANOVA entre ***: 15 puntos)*

Observación:
Hay variadas formas de efectuar los cálculos asociados a la descomposición de las sumas de cuadrados (para ello está la calculadora). Una de manera se resume a continuación:

Cuadro de totales y medias:

	Profesionales	*Operarios*	T_{i+}	$\overline{p}_{i+}$
Salarios	*12*	*11*	*23*	*3,833*
Comisiones	*22*	*18*	*40*	*6,667*
Participaciones	*25*	*21*	*46*	*7,667*
Bonificaciones	*16*	*12*	*28*	*4,667*
T_{+j}	*75*	*62*	*137*	
$\overline{p}_{+j}$	*6,250*	*5,167*		*5,708*

$$SC\_T = \sum_i \sum_j \sum_k p_{ijk}^2 - \frac{T^2}{n} = 857 - \frac{137^2}{24} = 857 - 782{,}042 = 74{,}958$$

$$SC\_A = \frac{\sum_i T_{i+}^2}{JK} - \frac{T^2}{n} = \frac{23^2 + 40^2 + 46^2 + 28^2}{(3)(2)} - 782{,}04 = 56{,}125$$

$$SC\_B = \frac{\sum_j T_{+j}^2}{IK} - \frac{T^2}{n} = \frac{75^2 + 62^2}{(3)(4)} - 782{,}042 = 7{,}042$$

$$SC\_E = \sum_i \sum_j \sum_k p_{ijk}^2 - \frac{\sum_i \sum_j T_{ij}^2}{K} = 857 - \frac{12^2 + 22^2 + 25^2 + 16^2 + 11^2 + 18^2 + 21^2 + 12^2}{3} = 10{,}667$$

$SC\_C = (SC\_T - SC\_A - SC\_B - SC\_E) = 1{,}125$

*****Por último, para evaluar la hipótesis que interesa se requiere revisar los supuestos,*

Independencia: Las 8 (JK) muestras de tamaño n = 3 son aleatorias
Normalidad: Las 8 (JK) poblaciones de donde se extraen las 8 (JK) muestras son normales
Homocedasticidad: Las 8 (JK) poblaciones tienen, todas ellas, la misma varianza

En el enunciado sólo se menciona que se tiene la Independencia, por lo cual será necesario suponer sin más que los otros supuestos se dan. Luego, se realiza el test::

Sobre el factor A (Formas de pago): Rechazar H_0 si: $F_A > {}_{1-\alpha}F_{I-1,\,n-IJ} = {}_{0{,}99}F_{3,\,16} = 5{,}29$. Puesto que 28,06 > 5,29, se rechaza H0 y se concluye que no todas las medias de productividad son iguales bajo las 4 formas de pago. Es decir, la forma de pago ejerce un efecto sobre la productividad.****

*(Conclusión entre ****: 5 puntos).*

Lo anterior es suficiente para la respuesta. Lo siguiente es complementario:

Sobre el factor B (Tipo de trabajador): Rechazar H_0 si $F_B > {}_{1-\alpha}F_{J-1,\,n-IJ} = {}_{0{,}99}F_{1,\,16} = 8{,}53$. Puesto que 10,56 > 8,53 se rechaza H_0 y se concluye que las medias de productividad difieren en profesionales y operarios. Es decir, el tipo de trabajador ejerce un efecto sobre la productividad.

Sobre la interacción: Rechazar H_0 si $F_C > {}_{1-\alpha}F_{(I-1)(J-1),\,n-IJ} = {}_{0{,}99}F_{3,\,16} = 5{,}29$. Puesto que 0,56 < 5,29, se mantiene H_0 y se concluye que la interacción entre el factor forma de pago y el factor tipo de trabajador no afecta a la productividad.

b) (15 ptos) ¿Qué forma de pago recomendarías para un profesional que se requiera contratar?

Solución
*** Se tiene las media maestrales según forma de pago:*

	Salarios	*Comisiones*	*Participaciones*	*Bonificaciones*
$\bar{p}_{i+}$	*3,833*	*6,667*	*7,667*	*4,667*

Aparentemente la forma de pago que impacta más en la productividad es la de "participación en utilidades" pero hay que someterlo a un test. De la indicación, se sigue que la idea es aplicar el criterio de Tukey. Para ello, se hacen las comparaciones de medias:

	$\bar{p}_{2+}$	$\bar{p}_{3+}$	$\bar{p}_{4+}$
$\bar{p}_{1+}$	$\lvert\bar{p}_{1+} - \bar{p}_{2+}\rvert = \lvert 3{,}83 - 6{,}67\rvert = 2{,}84$	$\lvert\bar{p}_{1+} - \bar{p}_{3+}\rvert = \lvert 3{,}83 - 7{,}67\rvert = 3{,}84$	$\lvert\bar{p}_{1+} - \bar{p}_{4+}\rvert = \lvert 3{,}83 - 4{,}67\rvert = 0{,}84$
$\bar{p}_{2+}$		$\lvert\bar{p}_{2+} - \bar{p}_{3+}\rvert = \lvert 6{,}67 - 7{,}67\rvert = 1$	$\lvert\bar{p}_{2+} - \bar{p}_{4+}\rvert = \lvert 6{,}67 - 4{,}67\rvert = 2$
$\bar{p}_{3+}$			$\lvert\bar{p}_{3+} - \bar{p}_{4+}\rvert = \lvert 7{,}67 - 4{,}67\rvert = 3$

Por otro lado: $DMS_{\text{Tukey}} = {}_{1-\alpha}q_{J,\,n-IJ}\sqrt{\frac{MCE}{r}} = {}_{0{,}99}q_{4,16}\sqrt{\frac{0{,}67}{6}} = 5{,}19\sqrt{0{,}112} = 1{,}73$

****
*(Cálculos entre **: 10 puntos)*

Luego, se observa que existen diferencias significativas para todos los pares de medias excepto para:

- ***Formas 1 y 4: (Salarios y Bonificaciones)***
- ***Formas 2 y3: (Comisiones y Participaciones)***

Por tanto, habría que recomendar vía comisiones o bien vía participaciones.***
*(Conclusión entre ***: 5 puntos)*

P3. En su famoso trabajo "*The use of multiple measurements in taxonomic problems"* (1936), R.A. Fisher utiliza cuatro variables asociadas a las flores de 3 especies: *Iris setosa, Iris versicolor* e *Iris viginica, con* muestras de 50 plantas de cada especie. Las variables son largo(L) y ancho(W) de los pétalos y los sépalos de las flores.

a) (3 ptos.) ¿Cuál es el número máximo que funciones discriminantes que podría calcular con esos datos?
Solución:
El número máximo de funciones discriminantes corresponde al número de clases de la variable de clasificación, menos 1. En este caso se tienen 3 clases (especies de plantas), luego se tendrán a lo más 2 funciones discriminantes.

b) (3 ptos.) ¿Qué puede decir acerca del poder discriminante del plano (Z_1, Z_2)?
Solución:
El cuadro de los Λ_q indica que Z_2 tiene un escaso poder discriminante, y que Z_1 tiene el mejor rendimiento discriminante. Esto se reafirma en el cuadro de valores propios, donde se observa la altísima correlación entre Z_1 y la clasificación (la primera correlación canónica)). Luego, el poder discriminante del plano (Z_1, Z_2) se debe a Z_1 y Z_2 no resulta interesante.

c) (3 ptos.) ¿Cuáles variables caracterizan a las funciones discriminantes?
Solución:
La matriz de estructura, que muestra las correlaciones de cada variable con las funciones discriminantes, permite ver la importancia relativa de cada variable en cada dimensión (función discriminante):

- ***El largo y ancho de los pétalos caracterizan mejor a Z_1***
- ***Ancho de pétalos y sépalos caracterizan mejor a Z_2***

d) (3 ptos.) Escriba las dos funciones discriminantes en su forma estandarizada.
Solución:
El cuadro de coeficientes estandarizados de las funciones discriminantes canónicas permite escribir las funciones:

$$Z^*_1 = -0.427 Sepal\_l^* - 0.521 Sepal\_w^* + 0.947 Petal\_l* + 0.575 Petal\_w*$$

$$Z^*_2 = 0.012 Sepal\_l^* + 0.735 Sepal\_w^* - 0.401 Petal\_l* + 0.581 Petal\_w*$$

donde * indica las variables "estandarizadas".

Observación: Los coeficientes estandarizados no conducen a las fórmulas directas de Z_1 y Z_2

e) (3 ptos.) ¿Cómo usaría las funciones discriminantes para asignar la especie a una nueva planta?
Solución:
Para hacer la clasificación de una nueva planta a partir de las funciones discriminantes, conviene utilizar los coeficientes no estandarizados y con ellos y los datos, generar las "puntuaciones" discriminantes. Con estas puntuaciones, y dado que en este caso sólo Z_1 es discriminante, se determinan los puntos de corte de los valores de Z_1 que discriminan las 3 especies. En nuevo casos e clasifica según indiquen esos puntos de corte.

Observación: No se presenta cuadros de coeficientes no estandarizados.

f) (4 ptos.) ¿Por qué los valores de la matriz de estructura difieren de los coeficientes de las funciones discriminantes canónicas estandarizadas?
Solución:
Porque los primeros corresponden a coeficientes totales** (no parciales), **similares a coeficientes de correlación, que reflejan la asociación** (no controlada) **de las funciones discriminantes y la variable**. En cambio, **los segundos

corresponden a contribuciones parciales *(asociación controlada)* ***de cada variable con las funciones discriminantes, cuando se controla por las demás variables independientes en el modelo****.*

g) (3 ptos.) ¿Cuáles especies, si las hay, son separadas por las dos funciones discriminantes?
Solución:
Del cuadro de los centroides de los grupos, se observa que:
- ***Z_1 separa Iris setosa de Iris versicolor e Iris virginica***
- ***Z_2 separa Iris versicolor de Iris setosa e Iris virginica***

h) (3 ptos.) ¿Cuántos casos fueron correctamente clasificados?
Solución:
Del cuadro de resultados de clasificación, se obtiene que 147 casos fueron bien clasificados por el modelo.