

SOLUCIÓN CONTROL N° 1

15/09/2006

1. a) Considere dos individuos u objetos caracterizados por variables no negativas $\mathbf{x}_r = (x_{1r}, x_{2r}, \dots, x_{pr})^t$ y

$$\mathbf{x}_s = (x_{1s}, x_{2s}, \dots, x_{ps})^t.$$

- i. Demuestre que la siguiente medida entre los objetos: $d_{rs} = 1 - \frac{2 \sum_{i=1}^p \min(x_{ir}, x_{is})}{\sum_{i=1}^p (x_{ir} + x_{is})}$, cumple las condiciones necesarias para ser un coeficiente o medida de disimilaridad. (2.0 pto.)

Solución:

- i) $0 \leq d_{rs} \leq 1$. Se sigue de forma inmediata al ser las variables positivas y teniendo en cuenta que $0 \leq 2 \min(x_{ir}, x_{is}) \leq x_{ir} + x_{is}$, $i = 1, \dots, p$
 $d_{rs} = 0 \Leftrightarrow \mathbf{x}_r = \mathbf{x}_s$. Si $\mathbf{x}_r = \mathbf{x}_s$, entonces $2 \min(x_{ir}, x_{is}) = 2 \min(x_{ir}, x_{ir}) = 2x_{ir} = x_{ir} + x_{is}$, por lo que $\frac{2 \sum_{i=1}^p \min(x_{ir}, x_{is})}{\sum_{i=1}^p (x_{ir} + x_{is})} = 1$ y $d_{rs} = 0$. Recíprocamente, si $d_{rs} = 0$, se debe cumplir que $\frac{2 \sum_{i=1}^p \min(x_{ir}, x_{is})}{\sum_{i=1}^p (x_{ir} + x_{is})} = 1$ y esto se alcanza sólo si $2 \min(x_{ir}, x_{is}) = x_{ir} + x_{is}$, $i = 1, \dots, p$.
 $d_{rs} = d_{sr}$. Es inmediato porque tanto la función mínimo como la función suma son simétricas en sus argumentos.

A esto se debe agregar la desigualdad triangular, que es bastante más difícil. No se considerará para el puntaje en caso que no se haya podido probar.

- ii. Obtenga una expresión explícita en el caso que las p variables sean binarias (0-1). (Ind: utilice la notación de coeficientes de similaridad entre datos categóricos, tal que $a + b + c + d = p$) (1.0 pto.)

Solución:

- ii) Sean a el número de coincidencias en 0 y d las coincidencias en el valor 1. Se tiene entonces que $\sum_{i=1}^p (x_{ir} + x_{is}) = b + c + 2d$. Por otro lado, $\min\{1, 0\} = 0$, $\min\{0, 1\} = 0$, $\min\{0, 0\} = 0$ y $\min\{1, 1\} = 1$ por lo que $2 \sum_{i=1}^p \min(x_{ir}, x_{is}) = 2d$. Con todo esto se obtiene que

$$d_{rs} = \frac{b + c}{b + c + 2d}$$

- b) Los administradores del estadio de los Diablos Aleatorios desean conocer si sus socios futboleros está interesados en cambiar la hora de inicio del match dominical, de las 18 hrs. a las 11:00 hrs. La planilla se compone de 5000 socios, sin embargo no se cuenta con el presupuesto para hacer la consulta a esa población.

- i. ¿Cuál debería ser el tamaño de un M.A.S., si se considera un nivel de confianza del 95% y se exige un error muestral de 3%? (1.5 ptos) Obs: Nivel de Confianza del 95% $\Rightarrow z = 1.96$

Solución:

Si no recuerdan la expresión para el error muestral de la proporción p (en este caso la proporción de socios que apoya el cambio de horario) en un MAS., deberán deducirlo como en clases. Dependiendo de los supuestos, es posible que la

expresión que obtengan no considere el factor de corrección por población finita. Así, el intervalo de confianza de a nivel $(1 - \alpha)\%$ está dado, el caso de muestreo con reemplazo, por:

$$\left[\hat{p} - \varphi_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}, \hat{p} + \varphi_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} \right] \quad (1)$$

Pero en este caso, es más adecuado considerar muestreo sin reemplazo, donde el intervalo está dado por:

$$\left[\hat{p} - \varphi_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1} \left(1 - \frac{n}{N}\right)}, \hat{p} + \varphi_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1} \left(1 - \frac{n}{N}\right)} \right] \quad (2)$$

Se puede usar, en todo caso, la primera fórmula y estará bien (especialmente cuando N es muy grande). Sin embargo, esto afectará la respuesta en el punto iii)

En esta solución, usaremos la fórmula (2) que es la más adecuada, pues $N=5000$ no es muy grande. Se tiene, $\alpha = 0.05$, $N = 5.000$. y $z = \varphi_{1-\alpha/2} = 1.96$. Se requiere aplicar el supuesto de varianza máxima para el caso de la proporción, de manera que $\hat{p} = 0.5$. Considerando un error de 0,03 se tiene:

$$0,03 = \varphi_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1} \left(1 - \frac{n}{N}\right)} = 1,96 \sqrt{\frac{0,25}{n-1} \left(1 - \frac{n}{5000}\right)}$$

$$\text{despejando se obtiene: } n = \frac{\left(\frac{0,03}{1,96}\right)^2 + 0,25}{\left(\frac{0,03}{1,96}\right)^2 + \frac{0,25}{5000}} = 880 \quad (*)$$

- ii. ¿Cómo cambia su respuesta si el nivel de confianza fuese 90% ($z = 1.64$)? ¿Qué significa esto? (0.5 pts)

Solución:

$$\text{Reemplazando } 1,96 \text{ por } 1,64 \text{ en } (*), \text{ se obtiene } n = \frac{\left(\frac{0,03}{1,64}\right)^2 + 0,25}{\left(\frac{0,03}{1,64}\right)^2 + \frac{0,25}{5000}} = 651. \text{ Esto significa que si estamos dispuestos a}$$

equivocarnos un 5% más (en términos absolutos) en nuestra estimación del intervalo de confianza para p , entonces puedo escoger una muestra más pequeña.

- iii. Al 95%, ¿qué sucede si se exige un menor error muestral (digamos 1%)? (0.5 pts)

Solución:

$$\text{Reemplazando } 0,03 \text{ por } 0,01 \text{ en } (*), \text{ se obtiene } n = \frac{\left(\frac{0,01}{1,96}\right)^2 + 0,25}{\left(\frac{0,01}{1,96}\right)^2 + \frac{0,25}{5000}} = 3288, \text{ es decir que un nivel de precisión muy}$$

alto (1% de error) nos exige una muestra demasiado grande, tomando en consideración que la población es 5000.

- iv. ¿Qué sucede con el tamaño de la muestra, si la población cambia a 10000 personas? (0.5 pts)

Solución:

$$\text{Reemplazando } 5000 \text{ por } 10000 \text{ en } (*), \text{ se obtiene } n = \frac{\left(\frac{0,01}{1,96}\right)^2 + 0,25}{\left(\frac{0,01}{1,96}\right)^2 + \frac{0,25}{10000}} = 965, \text{ es decir que si la población fuese el}$$

doble, el tamaño muestral requerido para la misma precisión no es el doble, sino sólo un 10% más de casos ($965/880 - 1$).

(Nótese que si en (i) se hubiese escogido la fórmula (1), entonces en esta parte la respuesta debía ser cualitativa, o sea indicar que si la población aumenta al doble, el tamaño muestral no aumenta en esa proporción sino que en una proporción menor o bastante menor).

2. Las negociaciones entre el sindicato y la gerencia de una gran empresa minera se concentran en incentivos salariales asociados a la productividad del trabajador. La firma tiene 5 divisiones (plantas). En cada una, a los trabajadores se les paga por comisión, salario o un plan de bonificaciones. Suponga que tres trabajadores, seleccionados para cada planta, se trataron con métodos distintos de pago. Su producción diaria, en una medida de unidades convencionales, aparece en el cuadro siguiente:

División	Comisión	Salario	Bonificaciones
1	25	25	37
2	35	25	50
3	20	22	30
4	30	20	40
5	25	25	35

- a) ¿Qué necesita para evaluar la aplicabilidad de ANOVA al problema? ¿cómo haría esa evaluación estadísticamente? (0.5 pts)

Solución:

Se requiere evaluar la normalidad de la variable de resultado (productividad, en este caso), para lo cual se puede considerar la prueba no paramétrica de Kolmogorov-Smirnov.

Además, se requiere que en cada grupo del tratamiento y bloque se tenga homogeneidad de varianzas para la variable de resultado (hipótesis de Homoscedasticidad). Para verificar este supuesto se puede considerar la prueba de Levene.

Observación: Algunos autores plantean que la no normalidad no afecta seriamente al estadístico F, por lo que sería aplicable la técnica de análisis de la varianza aún en casos en que la hipótesis de normalidad es rechazada.

- b) Considerando un nivel del 5%, ¿Cuál sería el plan de pagos que sugeriría a la gerencia si el objetivo es maximizar la producción? (3.0 pts.)

Solución:

Si cada individuo es escogido aleatoriamente según Tipo de Pago, y la medida de producción individual sigue una distribución normal, entonces podemos ocupar ANOVA de una vía para determinar si las producciones medias por Tipo de Pago son significativamente distintas o no. Aquí suponemos que la División a la que pertenece el sujeto no “bloquea” el efecto del Tipo de Pago que queremos probar.

Se tiene $n = 15$ casos, $c = 3$ el número de tratamientos y $SCT = 985.6$, $SCT = 613.2$ y $SCE = 372.4$. Con ello,
 $CMT = \frac{SCT}{n-1} = 70.4$, $CMTR = \frac{SCTR}{c-1} = 306.6$ y $CME = \frac{SCE}{n-c} = 31.0$. Así, para el estadístico de interés se observa
 $F = \frac{CMTR}{CME} = \frac{306.6}{31.0} = 9.88$ y de la tabla se tiene un valor crítico al nivel del 5% de $F_{2,12} = 3.89 < F$ con lo cual se

rechaza hipótesis de igualdad de medias, es decir los Tipos de Pago afectan la producción individual. Aunque es posible considerar una prueba de diferencia entre pares de medias como la DMS, basta mencionar que las medias en cada tratamiento (Comisión = 27.0; Salario = 23.4; Bonificaciones = 38.4) indican que la sugerencia que arroja máxima producción es la tercera: Bonificaciones.

- c) Si se piensa que podría ser necesario bloquear las divisiones. ¿Cómo cambia su respuesta anterior? (2.5 pts.)

Solución:

Se plantea la necesidad de un ANOVA de dos vías. A los cálculos de SCT y SCTR anteriores se agrega $SCBL = 250.3$, una nueva $SCE = SCT - SCTR - SCBL = 122.1$ y considerar $r = 5$. Con ello, $CMBL = \frac{SCBL}{r - 1} = 83.4$ y $CME = \frac{SCE}{(r - 1)(c - 1)} = 20.4$. Así, los estadísticos de interés arrojan $F = \frac{CMTR}{CME} = \frac{306.6}{15.3} = 20.1$ y $F' = \frac{CMBL}{CME} = \frac{62.6}{15.3} = 4.1$. Este último valor permite probar si el bloqueo es efectivo. Se contrasta con el valor de la tabla $F_{4,8} = 3.84 < F'$. Como se rechaza la hipótesis nula, se concluye que las divisiones tienen un efecto sobre la producción individual. Además, como $F_{2,8} = 4.46 < 20.1$, se confirma lo encontrado en la parte b), de manera que la respuesta no cambia aún después de bloquear por División.

3. Conteste las siguientes preguntas breve y justificadamente. No escriba más de 5 líneas por cada una.

- a) Cuando se trata de personas, al pasar del diseño muestral en el papel al trabajo de campo, desde el punto de vista del muestreo ¿Qué es lo más importante que debería tener en cuenta? (1.0 pto.)

Respuesta:

Vimos que el diseño que permite cuantificar el error que se comete en la inferencia es aquel que se asocia a un muestreo aleatorio. Luego, lo más importante es que el procedimiento utilizado en el trabajo de campo busque replicar las condiciones que caracterizan un muestreo aleatorio en teoría: independencia e idéntica distribución. O sea, el procedimiento debe garantizar imparcialidad en la selección y que las condiciones en que se lleva a cabo sean las mismas de un caso a otro.

- b) Con respecto a los métodos jerárquicos de agrupación, ¿cuál es la diferencia entre los enlaces simple, medio y completo? (1.0 pto.)

Respuesta:

La diferencia radica en los casos (reales o virtuales) utilizados como representativos de los grupos, que se usan para caracterizar la distancia entre los grupos. Así, Enlace simple mide la proximidad entre dos grupos calculando la distancia entre sus objetos más próximos; Enlace completo mide la proximidad entre dos grupos calculando la distancia entre sus objetos más lejanos; y Enlace medio entre grupos mide la proximidad entre dos grupos calculando la media de las distancias entre objetos de ambos grupos.

- c) ¿Qué es un dendrograma y para qué sirve? (1.0 pto.)

Respuesta:

En análisis de clusters jerárquicos, un dendrograma, también llamado diagrama de árbol, muestra el tamaño relativo de los coeficientes de proximidad para los cuales los casos fueron combinados. A coeficientes de distancia más grandes (o coeficientes de similitud más pequeños), más distintos los objetos combinados en los clusters, los cual puede ser poco deseable. Casos con baja distancia (alta similitud) están cerca, de manera que aparecen ligados en el dendrograma con una línea corta desde la izquierda, indicando que son aglomerados en un cluster de alta similitud. Por otro lado, cuando la línea que une los casos está a la derecha del dendrograma, el agrupamiento ocurre con un alto coeficiente de distancia,

indicando que los casos fueron aglomerados a pesar que podrían ser disímiles. Una interpretación inversa se tiene en el caso de medidas o coeficientes de similaridad.

- d) En un análisis de varianza, ¿por qué es importante la hipótesis de homocedasticidad? (1.0 pto.)

Respuesta:

La homogeneidad de varianzas (homoscedasticidad) en los grupos es importante ya que ellas solamente son atribuibles al error. Si no se tiene esta hipótesis, la variación observada entre grupos no sólo podría ser explicada por las diferencias entre las medias, sino a eventuales comportamientos diferenciados del error.

- e) Se ha dicho que los métodos exploratorios y confirmatorios vistos en clase producen síntesis de información en algún sentido: ¿Qué síntesis produciría y cómo se caracteriza la pérdida de información en el análisis discriminante? (1.0 pto.)

Respuesta:

En el caso de AD canónico, a partir de un conjunto de variables continuas, se busca un número acotado de dimensiones (nuevas variables), que permitan discriminar (es decir, que se relacionen lo mejor posible con) los grupos o clases en que está particionada una población. En general, se trata de una o dos dimensiones, ya sea un índice o un mapa, que permiten reproducir la clasificación así como predecir la pertenencia a un grupo de un nuevo caso. Hay varias medidas para la pérdida de información en un AD. Visto desde la perspectiva de cuán bueno es un AD, se tiene el Lamda de Wilks o la matriz de ajuste, por ejemplo.

- f) Si a usted se le entrega una base de datos con variables observadas sobre una muestra, ¿cuándo se justifica el método K-Means? (1.0 pto.)

Respuesta:

Se justifica la aplicación de K-Means cuando se busca clasificar casos o variables, es decir generar agrupaciones de casos o variable, y los datos son continuos.