INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

IN3401 - ESTADÍSTICA PARA ECONOMÍA Y GESTIÓN
Prof: Marcelo Henríquez
Auxs: Paulina Céspedes – Sebastián Rojas

Control Nº 1
Otoño 2012

1. *(6 puntos)* Considere la tabla de contingencia de las variables discretas X e Y:

x_i \ y_i	1	2	3
0	3	0	1
3	0	4	2
5	1	1	6

a) Encuentre las funciones de probabilidad marginales de X e Y.
b) Calcule las medias y las varianzas de X e Y.
c) Calcule la covarianza de X e Y.

Solución:

a) Se pide, para X: P(X=0); P(X=3); P(X=5) y para Y: P(Y=1); P(Y=2); P(Y=3)

**/ De la tabla de contingencia, se obtiene:*

$$P(X=0)=\sum_{j=1}^{3}P(X=0,Y=y_j)=\frac{n_{11}+n_{12}+n_{13}}{n}=\frac{n_{1\cdot}}{n}=\frac{4}{18}=\frac{2}{9}$$

De modo similar:

$$P(X=3)==\frac{n_{2\cdot}}{n}=\frac{6}{18}=\frac{3}{9}$$

$$P(X=5)==\frac{n_{3\cdot}}{n}=\frac{8}{18}=\frac{4}{9}$$

*Es fácil observar que P(X=0) + P(X=3) + P(X=5) = 1. /** *(1,0 puntos)*

**/ Por otro lado:*

$$P(Y=1)=\sum_{i=1}^{3}P(X=x_i,Y=1)=\frac{n_{11}+n_{21}+n_{31}}{n}=\frac{n_{\cdot 1}}{n}=\frac{4}{18}$$

De modo similar:

$$P(Y=2)==\frac{n_{\cdot 2}}{n}=\frac{5}{18}$$

$$P(X=5)==\frac{n_{\cdot 3}}{n}=\frac{9}{18}$$

*También es fácil observar que P(Y=1) + P(Y=2) + P(Y=3) = 1. /** *(1,0 puntos).*

b) Las medias…

$\bar{X}=\sum_{k=1}^{3}x_kP(X=x_k)=0\cdot P(X=0)+3\cdot P(X=3)+5\cdot P(X=5)=3\cdot\frac{3}{9}+5\cdot\frac{4}{9}=\frac{29}{9}=3,\bar{2}$ *(0,5 puntos)*

$\bar{Y}=\sum_{k=1}^{3}y_kP(Y=y_k)=1\cdot P(Y=1)+2\cdot P(Y=2)+3\cdot P(Y=3)=1\cdot\frac{4}{18}+2\cdot\frac{5}{18}+3\cdot\frac{9}{18}=\frac{41}{18}=2,2\bar{7}$ *(0,5 puntos)*

Las varianzas…

$V(X)=\sum_{k=1}^{3}(x_k-\bar{X})^2\cdot P(X=x_k)=\sum_{k=1}^{3}{x_k}^2\cdot P(X=x_k)-\bar{X}^2=\left\{0^2\cdot\frac{2}{9}+3^2\cdot\frac{3}{9}+5^2\cdot\frac{4}{9}\right\}-\left(\frac{29}{9}\right)^2=\frac{127}{9}-\frac{841}{81}=\frac{302}{81}=3,7284$ *(0,5 puntos)*

$V(Y)=\sum_{k=1}^{3}(y_k-\bar{Y})^2\cdot P(Y=y_k)=\sum_{k=1}^{3}{y_k}^2\cdot P(Y=y_k)-\bar{Y}^2=\left\{1^2\cdot\frac{4}{18}+2^2\cdot\frac{5}{18}+3^2\cdot\frac{9}{18}\right\}-\left(\frac{41}{18}\right)^2=\frac{105}{18}-\frac{1681}{324}=\frac{209}{324}=0,6451$ *(0,5 puntos)*

c) La covarianza…

$$Cov(X,Y)=\sum_{i=1}^{3}\sum_{j=1}^{3}(x_i-\bar{X})(y_j-\bar{Y})P(X=x_i,Y=y_j)=\sum_{i=1}^{3}\sum_{j=1}^{3}x_iy_jP(X=x_i,Y=y_j)-\bar{X}\bar{Y}$$
$$=\left\{0\cdot1\cdot\frac{3}{18}+0\cdot2\cdot\frac{0}{18}+0\cdot3\cdot\frac{1}{18}\right\}+\left\{3\cdot1\cdot\frac{0}{18}+3\cdot2\cdot\frac{4}{18}+3\cdot3\cdot\frac{2}{18}\right\}+\left\{5\cdot1\cdot\frac{1}{18}+5\cdot2\cdot\frac{1}{18}+5\cdot3\cdot\frac{6}{18}\right\}-\frac{29}{9}\cdot\frac{41}{18}$$

$= \frac{42}{18} + \frac{105}{18} - \frac{1189}{162} = \frac{67}{81} = 0{,}8272$ *(2,0 puntos)*

2. *(6 puntos)* La Encuesta Casen 2009 aseguraba que el 65% de las familias del Gran Santiago no tenía conexión a internet en su hogar. El Ministerio de Transportes y Telecomunicaciones lanza la iniciativa HCI ("Hogar Conectado a Internet") utilizando a empresas como intermediarios. Una de las empresas colaboradoras decide financiar con $160.000 la compra de un computador con conexión a Internet, a las familias que carezcan de ella en una muestra de 1000 familias elegidas al azar. Considerando la muestra:

   a) ¿Cuál es la probabilidad de que más de 750 familias no tengan conexión a internet?
   b) Encuentre un intervalo de confianza al 95% para el gasto total de la empresa en la iniciativa.

Solución:

a) *El problema puede modelarse como una muestra aleatoria $X_1, X_2, \dots, X_{1000}$ de Bernoulli (p), donde p = 0,65 es la probabilidad poblacional (suponemos que la Casen 2009 es precisa) de no tener conexión a Internet en el hogar.*

O sea $X_i = \begin{cases} 1 \text{ si la familia no tiene conexióna Internet en el hogar} \\ 0 \text{ si la familia tiene conexióna Internet en el hogar} \end{cases}$

Luego, se pregunta por $P = P(\sum_{i=1}^{1000} X_i > 750)$. *(1,0 punto)*

Esto equivale a preguntar por $P(\hat{p} > \frac{750}{1000})$ *(*) donde* $\hat{p} = \frac{1}{1000}\sum_{i=1}^{1000} X_i$ *es la proporción muestral, cuya varianza es* $\sigma_{\hat{p}}^2 = \frac{p(1-p)}{1000} = \frac{0{,}65(1-0{,}65)}{1000} = 2{,}275 \cdot 10^{-4}$. *(1,0 punto)*

Luego, considerando $\sigma_{\hat{p}} = \sqrt{2{,}275 \cdot 10^{-4}} = 1{,}508 \cdot 10^{-2}$ *y estandarizando la proporción muestral en (*) resulta:*

$P\left(\frac{\hat{p}-0{,}65}{1{,}508 \cdot 10^{-2}} > \frac{0{,}75-0{,}65}{1{,}508 \cdot 10^{-2}}\right) = P(Z > 6{,}630)$

Aplicando el TLC, se tiene que Z ~ N(0,1) y entonces: $P(Z > 6{,}630) = 1 - P(Z \leq 6{,}630) = 1 - \Phi(6{,}630)$

De la tabla se deduce que $\Phi(6{,}630) \approx 1$, *de donde la probabilidad de que más de de 750 familias no tengan conexión a Internet es (prácticamente) 0. (1,0 punto)*

b) *Del TLC es directo que un intervalo de confianza al 95% para* $\hat{p}$ *es* $\left[p - 1{,}96 \cdot \sigma_{\hat{p}},\ p + 1{,}96 \cdot \sigma_{\hat{p}},\right]$. *(1,5 puntos)*

Del enunciado se tiene que p = 0,65 y de la parte a) se tiene que $\sigma_{\hat{p}} = 1{,}508 \cdot 10^{-2}$. *Luego, el intervalo anterior es* $[0{,}620; 0{,}680]$. *Ahora bien, el gasto total de la firma en la iniciativa se puede estimar como* $\hat{g}$ *= $160.000×1.000×*$\hat{p}$*, de donde el intervalo pedido es:* $[\$160.000 \times 1.000 \times 0{,}620; \$160.000 \times 1.000 \times 0{,}680]$ = $[\$99.200.000; \$108.800.000]$ *(1,5 puntos)*

Nota: Está claro que los límites del intervalo para el gasto total deben ser divisibles por $160.000.- ¿no es cierto?

3. *(6 puntos)* Una empresa está interesada en evaluar 3 modalidades de capacitación de su fuerza de venta. Sesenta nuevos empleados son distribuidos aleatoriamente en tres grupos de igual tamaño. El grupo 1 recibe instrucción on-line, el grupo 2 la recibe mediante manuales y tutoriales impresos y el grupo 3 recibe entrenamiento en terreno. Al finalizar el curso se les aplica una evaluación (examen). El cuadro siguiente muestra una salida de SPSS con los intervalos de confianza al 95% para la media muestral de puntajes de la evaluación en cada grupo, la cual se ha contrastado con un valor esperado de 70 puntos:

	Valor de prueba = 70					
					95% Intervalo de confianza para la diferencia	
Grupo	t	gl	Sig. (bilateral)	Diferencia de medias	Inferior	Superior
1 Evaluación Ventas	1,504	19	,149	3,56766	-1,3975	8,5328
2 Evaluación Ventas	-2,125	19	,047	-6,42023	-12,7424	-,0980
3 Evaluación Ventas	9,415	19	,000	9,27924	7,2165	11,3420

   a) ¿Por qué en los contrastes se considera un estadístico t-Student? Explique.
   b) ¿Cuáles cursos producen, en promedio, resultados significativamente mejores que lo esperado? ¿Cuáles cursos producen, en promedio, resultados significativamente peores que lo esperado? Fundamente.

c) Indique que significa el valor Sig. (bilateral) que se reporta.

Solución:

a) Si se considera la media muestral de puntajes como un estimador de μ, el cuadro reporta el estadístico $t = \frac{\bar{X}-70}{s_{\bar{X}}}$ *(1,0 puntos), donde X es la v.a. puntaje en la evaluación (examen). Se utiliza este estadístico porque:*

- *Se estima μ con muestras pequeñas (en este caso tres muestras con n = 20 cada una). (1,0 punto)*
- *La varianza poblacional es desconocida y se estima por la varianza muestral* S^2*, de donde la varianza de la media muestral es estimada por* $S_{\bar{X}}{}^2 = \frac{S^2}{\sqrt{n}}$*. (0,5 puntos)*

Finalmente, si las variables aleatorias "Puntajes en la evaluación" son i.i.d. normales, entonces el estadístico t descrito sigue una distribución t-Student. (0,5 puntos)

Una respuesta que resume lo anterior es:
Se considera el estadístico $t = \frac{\bar{X}-70}{s_{\bar{X}}}$ *para calcular los intervalos de confianza, porque se está suponiendo que la media muestral de diámetros es Normal y su varianza desconocida; por ende, en el caso de muestras pequeñas (20 < 30 en este caso en cada muestra), la media muestral "estandarizada" sigue una t-Student.*

b) La respuesta puede basarse en los intervalos de confianza reportados para la diferencia entre la evaluación (examen) en cada muestra y el valor esperado:

Los intervalos de extremos estrictamente positivos ("intervalo positivo") indican que el resultado en la evaluación, al 95% de confianza, es mejor que el esperado $(\bar{X} > 70)$*: es el caso del grupo 3 (1,0 puntos). Los intervalos de extremos estrictamente negativos ("intervalo negativo") indican que el resultado, al 95% de confianza, es peor de que lo esperado* $(\bar{X} < 70)$*: es el caso 2. (1,0 puntos)*

No es necesario explicitarlo, pero los intervalos que incluyen el 0 (igualdad con el valor esperado) indican resultado que no es ni significativamente (al 5%) mejor ni significativamente peor al valor esperado $(\bar{X} \approx 70)$*: es el caso del grupo 1.*

Una respuesta alternativa se basa en el signo de las realizaciones del estadístico t y su significación (p-value):

En este caso, t > 0 indica que el resultado en el grupo es mejor que lo esperado (grupo 1 y 3), y t < 0 indica que el resultado en el grupo es peor que lo esperado (grupo 2). Pero estas diferencias son significativas solemente en los casos que la significación es pequeña (menor a 5%, por ejemplo). Aceptando la bilateralidad del test resporado, se tiene que las diferencias son significativas en el caso de los grupos 1 y 2.

c) La Sig, (bilateral) o p-value (bilateral) significa:

La probabilidad de equivocarse al rechazar la hipótesis nula que el resultado de la prueba es el esperado (el test es $H_0: \mu = 70$ *vs.* $H_1: \mu \neq 70$*). (1,2 puntos)*

(Una interpretación alternativa es: la probabilidad de que la diferencia entre el resultado en la prueba en el grupo y el puntaje esperado de 70 puntos, se deba al azar.)

Lo "bilateral" indica que esa probabilidad se calcula considerando las posibilidades tanto de exceso como de defecto en la diferencia. (0,8 puntos)

4. Para medir la tasa de desempleo en cierta zona, un investigador realiza una M.A.S. de 100 hogares. Después de visitar la zona, el investigador no ha logrado hacer contacto con nadie en 27 de los hogares. Preocupado por el sesgo asociado a no respuesta en su medición, selecciona ahora una M.A.S. de 27 hogares y aplica este procedimiento reiteradamente hasta lograr contacto en 100 hogares. La tasa de desempleo obtenida en la muestra final de 100 hogares es 12,8%. Analice la veracidad o falsedad de las siguientes afirmaciones.

a) La tasa de desempleo estimada está probablemente sesgada hacia arriba, es decir es demasiado alta.
b) La alta tasa de respuesta indica que se trata de un muestreo bien diseñado.
c) Se puede asegurar que la muestra final de 100 hogares también es aleatoria.

Solución:

a) *En efecto, si la tasa de desempleo se calcula considerando la fórmula del estimador "proporción muestral" de desempleados de un MAS, entonces probablemente la estimación esté sesgada. Esto porque en el muestreo descrito en el enunciado, las observaciones de cada etapa tienen distintos pesos en la conformación de la muestra final de 100 casos, no son necesariamente equiprobables como en el caso de MAS estricto. (1,0 puntos)*

Además, aún cuando se considere una fórmula que combine adecuadamente las observaciones de cada etapa, el diseño puede introducir un sesgo en la selección si las entrevistas se realizan en el día. Eventualmente, la imposibilidad de hacer contacto se puede deber a que los moradores el hogar están en sus trabajos, y esto en cada etapa del procedimiento descrito. Luego, la tasa de desempleo puede estar sobrestimada. (1,0 puntos)

Nota: Los sesgos asociados al procedimiento de selección muestral se denominan "sesgos de selección".

b) *En general, esto es falso. Mientras un buen diseño muestral no asegura alta tasa de respuestas (por ejemplo, si se aplica un cuestionario con algunas preguntas "difíciles"), tampoco una alta tasa de respuesta indica un muestreo bien diseñado. El ejemplo enunciado muestra que un procedimiento por etapas, diseñado para alcanzar una alta tasa de respuesta, puede introducir sesgos. (2,0 puntos)*

c) *En general, las observaciones de desempleo en los hogares en la muestra final no constituyen una muestra aleatoria, en términos estrictos de la definición. En efecto, si bien son independientes, estas observaciones no son idénticamente distribuidas ya que la población (segmento de zona) donde se aplica cada etapa de muestreo no es la misma. (1,0 puntos)*

Sin embargo, si por "aleatoriedad" se entiende la introducción del azar en el procedimiento de selección, entonces cabe señalar que en la muestra final efectivamente esto se ha efectuado, pero no con la "total imparcialidad" que se requiere. (1,0 puntos).

$$NC1 = \frac{(P1 + P2 + P3 + P4)}{24} + 1$$