Pontificia Universidad Católica de Chile
Facultad de Matemáticas
Departamento de Estadística

Temporada Académica de Verano (enero 2024)

Curso	:	**Probabilidad y Estadística**
Sigla	:	**EYP1113**
Profesores	:	**Ricardo Aravena C., Felipe Ossa M.**

PAUTA INTERROGACIÓN 3

Pregunta 1

Considere que se observa una muestra i.i.d. $X_1, \ldots, X_n$ de una población con distribución Exponencial Trasladada$(\nu, 2)$, cuya función de densidad es

$$f_X(x) = \begin{cases} \nu\, e^{-\nu\,(x-2)} & \text{si } x > 2 \\ 0 & \text{si } x \leq 2 \end{cases}$$

y $\nu > 0$ es un parámetro desconocido.

(a) **[1 punto]** Encuentre $\hat{\nu}_{\text{MM}}$, el estimador de momentos de ν.

(b) **[2 puntos]** Encuentre $\hat{\nu}_{\text{MV}}$, el estimador de ν por máxima verosimilitud.

(c) **[1.5 puntos]** Calcule el sesgo de $\hat{\nu}_{\text{MV}}$ y compruebe que el estimador es asintóticamente insesgado para ν.

(d) **[1.5 puntos]** Calcule la varianza de $\hat{\nu}_{\text{MV}}$ y, junto con (c), verifique es un estimador consistente de ν.

Pista: Si $W \sim$ Gamma-Trasladada(α, ν, c), entonces

$$E\left(\frac{1}{W-c}\right) = \frac{\nu}{\alpha - 1}\,, \qquad E\left(\frac{1}{(W-c)^2}\right) = \frac{\nu^2}{(\alpha-1)(\alpha-2)}$$

Si $T_1, \ldots, T_n \sim$ Exp-Trasladada(ν, d), entonces $\sum T_i \sim$ Gamma-Trasladada$(n,\, \nu,\, nd)$.

Solución:

(a) La esperanza poblacional es $E(X) = \frac{1}{\nu} + 2$, por lo tanto $\hat{\nu}_{\text{MM}}$ satisface

$$\frac{1}{\hat{\nu}} + 2 = \bar{X}$$

Por lo tanto,

$$\hat{\nu}_{\text{MM}} = \frac{1}{\overline{X} - 2} = \frac{n}{\sum_{i=1}^n (X_i - 2)} = \frac{n}{\sum_{i=1}^n X_i - 2n}$$

(b) La función de verosimilitud de ν está dada por

$$\begin{aligned} L(\nu) &= \prod_{i=1}^n f(x_i) \\ &= \prod_{i=1}^n \nu\, \exp\left\{-\nu\,(x_i - 2)\right\} \\ &= \nu^n\, \exp\left\{-\nu \sum_{i=1}^n x_i + 2n\,\nu\right\} \end{aligned}$$

Para maximizar $L(\nu)$ en ν, se recomienda tomar el logaritmo de la verosimilitud de ν. Este es,

$$\ell(\nu) = \ln L(\nu) = n\,\ln\nu - \nu\sum_{i=1}^{n} x_i + 2\,n\,\nu$$

Derivando respecto a ν e igualando a cero, entonces $\hat{\nu}$ debe satisfacer,

$$\begin{aligned}
\frac{d}{d\nu}\ln L(\nu) &= \frac{n}{\hat{\nu}} - \sum_{i=1}^{n} x_i + 2n \\
0 &= \frac{n}{\hat{\nu}} - \sum_{i=1}^{n} x_i + 2n \\
\hat{\nu}_{\textsf{MV}} &= \frac{n}{\sum_{i=1}^{n} x_i - 2n}
\end{aligned}$$

Por lo tanto, el estimador máximo verosímil es

$$\hat{\nu}_{\textsf{MV}} = \frac{n}{\sum_{i=1}^{n} x_i - 2n} = \frac{n}{\sum_{i=1}^{n}(x_i - 2)} = \frac{1}{\overline{X} - 2}$$

(c) Usando la ayuda, notar que $W = \sum_{i=1}^{n} X_i$ tiene distribución Gamma-Trasladada$(n,\,\nu,\,2n)$. La esperanza de $\hat{\nu}$ es

$$E\left(\hat{\nu}\right) = E\left(\frac{n}{\sum X_i - 2n}\right) = n\,E\left(\frac{1}{W - 2n}\right) = \frac{n\,\nu}{n-1}$$

Alternativamente, $X_i - 2 \sim$ Exp(ν), por lo tanto $W' = \sum_{i=1}^{n}(X_i - 2) \sim$ Gamma(n, ν), y así,

$$E\left(\hat{\nu}\right) = n\,E\left(\frac{1}{W'}\right) = \frac{n\,\nu}{n-1}$$

El sesgo es

$$\text{sesgo} = E(\hat{\nu}) - \nu = \frac{n}{n-1}\,\nu - \nu = \frac{\nu}{n-1}$$

y converge a 0 cuando $n \to \infty$, por lo tanto $\hat{\nu}$ es asintóticamente insesgado. Si se demuestra que $E(\hat{\nu}) \to \nu$, dar también todo el puntaje.

(d) Usando argumento similar a (c) se obtiene la varianza de $\hat{\nu}$,

$$\begin{aligned}
\text{Var}(\hat{\nu}) &= \text{Var}\left(\frac{n}{W - 2n}\right) \\
&= n^2\left[E\left(\frac{1}{(W-2n)^2}\right) - E\left(\frac{1}{W-2n}\right)^2\right] \\
&= n^2\left[\frac{\nu^2}{(n-1)(n-2)} - \frac{\nu^2}{(n-1)^2}\right] \\
&= \frac{n^2\nu^2}{(n-1)^2(n-2)}
\end{aligned}$$

y converge a 0 cuando $n \to \infty$. Como tanto varianza y sesgo convergen a cero con $n \to \infty$ entonces se dice que $\hat{\nu}$ es un estimador consistente para ν.

Pregunta 2

En TAV es común que los y las estudiantes estudien en grupos. Usted busca comprobar que más del 40 % de los estudiantes lo hacen. Por otra parte, se dice respecto a los estudiantes *que estudian en forma solitaria*, que estos obtienen una nota media inferior a 5.0.

Con el objetivo de comprobar o refutar sus apreciaciones (hipótesis), aplica una encuesta a los $n = 113$ alumnos del curso TAV 2024, obteniendo los siguientes datos muestrales:

Característica	Respuesta
Número de encuestados	113
¿Estudia solo? $\to$ Sí	58
De los que estudian solos:	
Nota promedio	4.76
Desv. estándar	0.7

Lleve a cabo las dos pruebas de hipótesis correspondientes, use un nivel de significancia $\alpha = 1\,\%$ cada uno. Sea explícito: hipótesis, estadístico, criterio (valor-p o valor crítico) y conclusión en el contexto del problema. **[3 puntos] por cada prueba**.

Solución:

- *Más del 40 % de los estudiantes estudian en grupo.* Sea p la verdadera proporción de estudiantes que estudian en grupo. Las hipótesis a contrastar son

$$H_0 : p = 0.4 \qquad \text{vs} \qquad H_a : p > 0.4$$

Se rechaza H_0 si

$$Z_0 = \frac{\hat{p} - 0.4}{\sqrt{\frac{0.4 \cdot 0.6}{n}}} > k_{0.99}$$

donde $\hat{p} = \frac{113-58}{113} = 0.487$, $n = 113$ y $k_{0.99} = 2.33$.

Vemos que

$$Z_0 = \frac{0.487 - 0.4}{\sqrt{\frac{0.4 \cdot 0.6}{113}}} = 1.888$$

entonces como $1.888 \not> 2.33$, no se rechaza H_0, o bien no existe evidencia estadística suficiente para afirmar que más que el 40 % de los estudiantes estudia en grupos, con un 1 % de significancia.

Alternativamente, se puede calcular el valor-p de Z_0,

$$\text{valor-p} = P(Z > 1.888) = 1 - \Phi(1.89) = 0.0294$$

como valor-p $\not< \alpha$, no se rechaza H_0, o bien no existe evidencia estadística suficiente para afirmar que más que el 40 % de los estudiantes estudia en grupos, con un 1 % de significancia.

(Si se hizo test de hipótesis para la proporción de estudiantes que estudian en solitario $H_0 : p = 0.6$, $H_a : p < 0.6$, es consistente y se otorga puntaje equivalente)

- *Estudiantes que estudian en forma solitaria tienen nota media inferior a 5.0*. Sea μ la nota media en todos estos estudiantes que estudian en solitario. Las hipótesis a contrastar son

$$H_0 : \mu = 5 \qquad \text{vs} \qquad H_a : \mu < 5$$

Se rechaza H_0 si

$$Z_0 = \frac{\bar{X} - 5}{S/\sqrt{n}} < -t_{0.99;n-1}$$

donde $\bar{X} = 4.76$, $S = 0.7$, $n = 58$ y como $n > 30$ se puede utilizar simplemente percentil normal $t_{0.99} \approx 2.33$.

Calculamos el estadístico Z_0,

$$Z_0 = \frac{4.76 - 5}{0.7/\sqrt{58}} = -2.61$$

entonces como $-2.61 < -2.33$, se rechaza H_0 o bien existe evidencia suficiente para afirmar que de los estudiantes que estudian en solitario, la nota media es menor a 5.0, con un 1 % de significancia.

Alternativamente, se puede calcular el valor-p,

$$\text{valor-p} = P(Z < -2.61) = 1 - \Phi(2.61) = 1 - 0.9955 = 0.0045$$

entonces como $0.0045 < \alpha$, se rechaza H_0 o bien existe evidencia suficiente para afirmar que de los estudiantes que estudian en solitario, la nota media es menor a 5.0, con un 1 % de significancia.

Pregunta 3

Un estudio respecto a las horas de estudio diario en TAV, según género, es llevado a cabo. Parte de los resultados muestrales se presentan en la siguiente tabla (asuma que los tiempos son independientes y se comportan de acuerdo con una distribución normal):

Característica	Masculino	Femenino	Total
Número de casos	16	14	30
Tiempo: (en min.)			
Promedio	46	52	49
Desv. estándar	13	8	11

(a) **[2 puntos]** Considerando la muestra completa, ¿existe evidencia que permita afirmar que la variabilidad (en términos de la desv.estándar) de las horas de estudio es inferior a 15 min? Use $\alpha = 5\,\%$. Sea explícito: hipótesis, estadístico, criterio (valor-p o valor crítico) y conclusión en el contexto del problema.

(b) **[4 puntos]** ¿Existe evidencia que permita afirmar que el género Femenino dedica un mayor tiempo medio al estudio que sus congéneres Masculinos? Use $\alpha = 10\,\%$ y asuma varianzas iguales. Sea explícito: hipótesis, estadístico, criterio (valor-p o valor crítico) y conclusión en el contexto del problema.

Solución:

(a) Las hipótesis a contrastar son

$$H_0 : \sigma = 15 \qquad , \qquad H_a : \sigma < 15$$

Por lo tanto un test de hipótesis apropiado rechaza la hipótesis nula cuando

$$C_0 = \frac{(n-1)S^2}{\sigma_0^2} < c_{0.05;n-1}$$

donde $c_{0.05;29}$ es el percentil 5 % de la distribución chi cuadrado $\chi^2(29)$. Por tabla, $c_{0.05;29} = 17.708$.

El valor del estadístico es

$$C_0 = \frac{29 \cdot 11^2}{15^2} = 15.6$$

y entonces, como $15.6 < 17.71$, se rechaza H_0. Es decir, existe evidencia suficiente para afirmar que la desviación estándar de tiempo dedicado al estudio es menor a 15 minutos, con un 5 % de significancia.

Alternativamente, se puede ver utilizando tabla $\chi^2(29)$ que

$$c_{0.005} < 15.6 < c_{0.025}$$

por lo tanto valor-p de estadístico C_0 se encuentra entre 0.5 % y 2.5 %. Como es menor que α, se concluye coherentemene de la misma manera como en párrafo anterior.

(b) Las hipótesis a contrastar son

$$H_0 : \mu_F = \mu_M \qquad , \qquad H_a : \mu_F > \mu_M$$

Un test de hipótesis apropiado rechaza la hipótesis nula H_0 si

$$T_0 = \frac{\bar{T}_F - \bar{T}_M}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} > t_{0.9;\, n+m-2}$$

donde $t_{0.9;\, n+m-2}$ corresponde al percentil 90 % de distribución t-student(28), y S_p es la estimación de desviación estándar combinada,

$$S_p = \sqrt{\frac{1}{28}(15 \cdot 13^2 + 13 \cdot 8^2)} = 10.966$$

Por tabla obtenemos $t_{0.9;\, 28} = 1.313$. Por lo tanto, el valor del estadístico del test es

$$T_0 = \frac{52 - 46}{10.966\sqrt{\frac{1}{16} + \frac{1}{14}}} = 1.495$$

Como $1.495 > 1.313$ se rechaza la hipótesis nula H_0, y podemos afirmar, con 10 % de significancia, que el género femenino dedica en promedio más tiempo al estudio que el género masculino en este curso.

Alternativamente, como

$$t_{0.9} = 1.313 < 1.495 < 1.701 = t_{0.95}$$

entonces valor-p se encuentra entre 5 % y 10 %, y como es menor que α, igualmente que párrafo anterior se rechaza H_0.

(Si se calculó el estadístico con $\bar{T}_M - \bar{T}_F$ en el numerador, entonces la región crítica y valor de T_0 es completamente simétrico y la conclusión debe ser exactamente la misma: $T_0 = -1.495$ y $t_{0.1;28} = -1.313$)

Pregunta 4

Se plantea un modelo de regresión lineal simple para explicar el comportamiento esperado de una variable Y como función de X.

$$E(Y|X = x) = \beta_0 + \beta_1\, x$$

Asuma que todos los supuestos de dicho modelo se cumplen. Se entrega la siguiente información resumen a partir de 26 datos observados (x_i, y_i),

$$\sum_{i=1}^{26}(x_i - \bar{x})^2 = 24.7 \quad , \qquad \sum_{i=1}^{26}(y_i - \bar{y})^2 = 936.3\,, \qquad \sum_{i=1}^{26}(x_i - \bar{x})(y_i - \bar{y}) = -131.7$$

$$\bar{X} = \frac{1}{26}\sum_{i=1}^{26} x_i = 3.35 \quad , \qquad \bar{Y} = \frac{1}{26}\sum_{i=1}^{26} y_i = 19.8$$

(a) **[1.5 puntos]** Calcule las estimaciones de pendiente β_1 e intercepto β_0 de la recta de regresión del modelo.

(b) **[1 punto]** Demuestre que, en general,

$$SCR = \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1\, x_i - \bar{y})^2 = (\hat{\beta}_1)^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$$

(c) **[1.5 puntos]** Calcule la estimación de σ^2 del modelo (**Pista:** recuerde que $s^2_{Y|x} = \frac{SCE}{n-2}$)

(d) **[2 puntos]** Compruebe la existencia de la regresión con test de hipótesis para β_1. Con los datos entregados, ¿se puede concluir que la pendiente es significativamente diferente de cero? Ocupe un $\alpha = 5\,\%$ de significancia.

Pista: recordar fórmula de error estándar de $\hat{\beta}_1$

$$s_{\hat{\beta}_1} = \frac{s_{Y|x}}{\sqrt{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Solución:

(a) Las estimaciones de β_1 y β_0 son

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{26}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{26}(x_i - \bar{x})^2} = \frac{-131.7}{24.7} = -5.332$$

y

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \cdot \overline{X} = 19.8 + 5.332 \cdot 3.35 = 37.66$$

(b) Por definición, remplazamos $\hat{\beta}_0$ y

$$\begin{aligned} SCR &= \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1\, x_i - \bar{y})^2 \\ &= \sum_{i=1}^{n}(\bar{y} - \hat{\beta}_1\, \bar{x} + \hat{\beta}_1\, x_i - \bar{y})^2 \\ &= \sum_{i=1}^{n}(-\hat{\beta}_1\, \bar{x} + \hat{\beta}_1\, x_i)^2 \\ &= \sum_{i=1}^{n}[\hat{\beta}_1\,(x_i - \bar{x})]^2 \\ &= (\hat{\beta}_1)^2\, \sum_{i=1}^{n}(x_i - \bar{x})^2 \end{aligned}$$

(c) Utilizando (b) tenemos que

$$SCE = SCT - SCR = \sum_{i=1}^{26}(y_i - \bar{y})^2 - (\hat{\beta}_1)^2\, \sum_{i=1}^{26}(x_i - \bar{x})^2 = 936.3 - (-5.332)^2 \cdot 24.7 = 234.07$$

Por lo tanto

$$s^2_{Y|x} = \frac{SCE}{n-2} = \frac{234.07}{24} = 9.752$$

(d) Las hipótesis a contrastar son

$$H_0 : \beta_1 = 0 \qquad \text{vs} \qquad \beta_1 \neq 0$$

El test de hipótesis correspondiente rechaza la hipótesis nula H_0 si

$$|T_0| = \frac{|\hat{\beta}_1 - 0|}{s_{\hat{\beta}_1}} > t_{0.975\,;\,n-2}$$

Por lo tanto, calculamos el error estándar de $\hat{\beta}_1$,

$$s_{\hat{\beta}_1} = \frac{s_{Y|x}}{\sqrt{\sum_{i-1}^{n}(x_i - \bar{x})^2}} = \frac{\sqrt{9.752}}{\sqrt{24.7}} = 0.628$$

Luego,

$$T_0 = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{-5.332}{0.628} = -8.49$$

mientras que $t_{0.975;\,24} = 2.063$. Como $|-8.49| > 2.063$, entonces con un 5 % de significancia se concluye que la pendiente es distinta de cero, y por lo tanto se dice que existe regresión lineal entre Y y X.

Tiempo: 2 horas