Pontificia Universidad Católica de Chile
Facultad de Matemáticas
Departamento de Estadística

Primer Semestre 2023

Curso	:	**Probabilidad y Estadística**
Sigla	:	**EYP1113**
Profesores	:	**Ricardo Aravena C., Ricardo Olea O.,**
		Felipe Ossa M. y Alejandro Sepulveda P.

PAUTA INTERROGACIÓN 4

Pregunta 1

Numerosos antecedentes ponen entredichos la acción de las fundaciones que son, parcialmente, financiadas por el estado. Usted, interesado en dilucidar ciertos aspectos relacionados con las fundaciones lleva a cabo un estudio respecto a: Antigüedad (antes 2021.09 y desde 2021.09) y montos traspasados (en mm CL$).

Por lo anterior obtiene una muestra de tamaño 52, registrando los diversos antecedentes que son presentados en la tabla adjunta:

	antes 2021.09	desde 2021.09	total
N° de fundaciones	29.0	23.0	52.0
Montos			
mean	37.4	40.8	38.9
sd	6.0	8.2	7.0

(a) **[2.0 Ptos.]** ¿Existe evidencia que más de un tercio de las fundaciones se crearon el 2021.09 o posterior? Utilice un nivel de significancia del $5\,\%$.

(b) **[3.0 Ptos.]** ¿Es válido afirmar que las fundaciones de creación reciente reciben montos medios superiores a las fundaciones con existencia anterior? Asuma Normalidad y utilice un nivel de significancia del $10\,\%$.

(c) **[1.0 Ptos.]** Con base a la información obtenida, se desea replicar el estudio, ¿cuántas fundaciones deberían ser seleccionadas a fin de estimar, con un 95 % de confianza, los montos medios totales con un error menor o igual a 1.5 mm CL$?

mm CL$: Millones de pesos chilenos.

Solución

(a) Sea p la proporción de fundaciones creadas el 2021.09 o posterior.

Se pide contrastar las siguientes hipótesis:

$$H_0 : p = p_0 \quad \text{vs} \quad H_a : p > p_0,$$

con $p_0 = 1/3$.

Bajo H_0 se tiene que

$$Z_0 = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0 \cdot (1 - p_0)}{n}}} \overset{\text{aprox}}{\sim} \text{Normal}(0, 1)$$

Del enunciado se tiene que $\hat{p} = \dfrac{23}{52}$ y reemplazando en Z_0 se tiene que:

$$Z_0 = 1.666987 \rightarrow \text{valor-p} \approx 1 - \Phi(1.67) = 0.0475$$

Por lo tanto, considerando un nivel de significancia del 5 %, se rechaza H$_0$ y se apoya la afirmación que más de un tercio de las fundaciones se crearon el 2021.09 o posterior.

Alternativamente en vez de calcular valor-p, se podría obtener el valor critico $k_{0.95} \approx 1.645$ y como $Z_0 > k_{0.95}$, entonces se rechaza H$_0$ y se apoya la afirmación que más de un tercio de las fundaciones se crearon el 2021.09 o posterior.

(b) Definamos como X a los montos antes de 2021.09 y como Y a los montos desde 2021.09.

Se pide contrastar las siguientes hipótesis:

$$H_0 : \mu_X = \mu_Y \quad \text{vs} \quad H_a : \mu_X < \mu_Y,$$

Bajo el supuesto que las varianzas son iguales se tiene que:

$$F_0 = \frac{S_X^2}{S_Y^2} = 0.5353956 > F_{1-0.10/2}(29-1, 23-1) = \frac{1}{F_{0.95}(22, 28)} = 0.5181347$$

o

$$F_0 = \frac{S_Y^2}{S_X^2} = 1.867778 < F_{1-0.10/2}(29-1, 23-1) = F_{0.95}(28, 22) = 2.00.$$

Por lo tanto el estadístico de prueba bajo H$_0$ que se utilizará, asumirá varianzas desconocidas, pero iguales:

$$T_0 = \frac{\overline{X} - \overline{Y}}{S_p \sqrt{\frac{1}{n} + \frac{1}{n}}} \sim \text{t-Student}(n + m - 2)$$

con $S_p = \frac{(n-1)\, S_X^2 + (m-1)\, S_Y^2}{n+m-2}$.
Reemplazando $n = 29$, $m = 23$, $\overline{X} = 37.4$, $\overline{Y} = 40.8$, $S_X = 6$ y $S_y = 8.2$, se tiene que $T_0 = -1.726485$.

Como $n + m - 2 > 30$, entonces el valor-p $\approx 1 - \Phi(1.73) = 0.0418$ y el valor critico $t_{0.10/2}(n + m - 2) \approx k_{0.05} = -1.645$.

Por lo que es posible afirmar que las fundaciones de creación reciente reciben montos medios superiores a las fundaciones con existencia anterior.

(c) Para el tamaño de muestra, el valor de σ será el obtenido en este estudio:

$$n = \left(\frac{k_{0.975} \cdot \sigma}{\text{e.e}}\right)^2 = \left(\frac{1.96 \cdot 7}{1.5}\right)^2 = 83.66 \rightarrow \text{84 casos}$$

Asignación de puntaje:

Logro 1: Definir correctamente H$_a$ [0.2 Ptos.] y obtener Z_0. [0.8 Ptos.]

Logro 2: Calcular correctamente el valor-p o determinar el valor crítico. [0.8 Ptos.]. Concluir correctamente. [0.2 Ptos.]

Logro 3: Plantear correctamente las hipótesis [0.5 Ptos.] y determinar correctamente que el supuesto de varianzas iguales no se puede rechazar [0.5 Ptos.].

Logro 4: Calcular correctamente T_0. [1.0 Ptos.]

Logro 5: Calcular valor-p o valor critico [0.5 Ptos.] y concluir correctamente [0.5 Ptos.].

Logro 6: Determinar correctamente el tamaño muestral de la replica de este estudio. [1.0 Ptos.]

+ 1 Punto Base

Pregunta 2

Ayer miércoles 5 de julio a las 20:31:56 horas ocurrió un sismo de 5.8° en la escala de Richter. Las magnitudes de los n sismos que le siguieron distribuyen Log-Normal(λ, ζ). Bajo el supuesto que las magnitudes se comportan de manera independiente y que el parámetro ζ es conocido:

(a) **[1.0 Ptos.]** Obtenga el estimador de momento de λ.

(b) **[3.0 Ptos.]** Obtenga el estimador de máxima verosimilitud de λ y su distribución asintótica.

(c) **[2.0 Ptos.]** A partir del ECM aproximado de 1er orden compruebe que el estimador de momento es consistente, pero menos eficiente que el EMV.

Solución

Definamos como $X_1, \ldots, X_n$ las magnitudes de los n sismos posteriores, que según enunciado son variables aleatorias iid Log-Normal(λ, ζ), con ζ conocido.

Del formulario se tiene:

$$\mu_X = e^{\lambda+\zeta^2/2}, \quad \sigma_X^2 = \mu_X^2\,(e^{\zeta^2}-1) \quad \text{y} \quad \mathsf{E}(X^k) = e^{k\,\lambda+(k\,\zeta)^2/2}.$$

(a) Igualando 1er momento teórico al empírico se obtiene el EM de λ:

$$\mathsf{E}(X) = e^{\lambda+\zeta^2/2} = \overline{X} \to \widetilde{\lambda} = \ln(\overline{X}) - \frac{\zeta^2}{2}.$$

(b) La Verosimilitud y Log Verosimilitud está dada por

$$L(\lambda) = (2\,\pi)^{-n/2}\cdot\zeta^{-n}\cdot\left(\prod_{i=1}^n X_i\right)^{-1}\cdot\exp\left\{-\frac{1}{2\,\zeta^2}\sum_{i=1}^n(\ln(X_i)-\lambda)^2\right\}, \tag{1}$$

$$\ln\,L(\lambda) = -\frac{n}{2}\,\ln(2\,\pi) - n\,\log(\zeta) - \sum_{i=1}^n \ln(X_i) - \frac{1}{2\,\zeta^2}\sum_{i=1}^n(\ln(X_i)-\lambda)^2. \tag{2}$$

Derivando parcialmente (2) con respecto a λ y luego igualando a cero se obtiene el EMV de λ:

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^n \ln(X_i).$$

A partir de la 2da derivada parcial se obtiene la información de Fisher y su CCR:

$$I(\lambda) = -\mathsf{E}\left(\frac{\partial^2}{\partial\,\lambda^2}\ln\,L(\lambda)\right) = \frac{n}{\zeta^2} \to \text{CCR} = \frac{\zeta^2}{n}.$$

Por lo tanto se tiene que

$$\hat{\lambda} \overset{\text{aprox}}{\sim} \text{Normal}\left(\lambda,\, \frac{\zeta}{\sqrt{n}}\right).$$

Como $\ln(X_i) \overset{\text{iid}}{\sim}$ Normal(λ, ζ), además se puede afirmar que

$$\hat{\lambda} \overset{\text{exacta}}{\sim} \text{Normal}\left(\lambda,\, \frac{\zeta}{\sqrt{n}}\right).$$

(c) Aplicando aproximación por método delta de 1er orden para $\widetilde{\lambda}$, se tiene que:

$$\begin{aligned}
\widetilde{\lambda} = g(X_1, \ldots, X_n) &\overset{\text{id}}{\approx} \left[\ln\left(\frac{n\cdot\mu_X}{n}\right) - \frac{\zeta^2}{2}\right] + \sum_{i=1}^{n} \frac{1/n}{\left(\frac{n\cdot\mu_X}{n}\right)}\,(X_i - \mu_X)\,, \\
&= \left[\ln\left(\frac{n\cdot e^{\lambda+\zeta^2/2}}{n}\right) - \frac{\zeta^2}{2}\right] + \sum_{i=1}^{n} \frac{1}{n\cdot\mu_X}\,(X_i - \mu_X)\,, \\
&= \lambda + \sum_{i=1}^{n} \frac{1}{n\cdot\mu_X}\,(X_i - \mu_X)\,.
\end{aligned}$$

Aplicando E$(\cdot)$ y Var$(\cdot)$:

$$\text{E}(\widetilde{\lambda}) = \lambda \quad \text{y} \quad \text{Var}(\widetilde{\lambda}) \overset{\text{ind}}{=} \frac{n\cdot\sigma_X^2}{n^2\cdot\mu_X^2} = \frac{e^{\zeta^2}-1}{n}$$

Como

$$\text{ECM}(\widetilde{\lambda}) = \frac{e^{\zeta^2}-1}{n} \to 0,$$

cuando $n\to\infty$, entonces $\widetilde{\lambda}$ es un estimador consistente para λ.

Finalmente

$$\frac{\text{ECM}(\widetilde{\lambda})}{\text{ECM}(\hat{\lambda})} = \frac{e^{\zeta^2}-1}{\zeta^2} > 1,$$

debido a que $\lim_{\zeta\to 0} \frac{e^{\zeta^2}-1}{\zeta^2} \overset{L'H}{=} 1$ y $\lim_{\zeta\to+\infty} \frac{e^{\zeta^2}-1}{\zeta^2} \overset{L'H}{=} +\infty$.

Por lo tanto el EMV es más eficiente que el EM.

Asignación de puntaje:

Logro 1: Obtener EM. [1.0 Ptos.]

Logro 2: Obtener EMV. [1.0 Ptos.]

Logro 3: Obtener $I(\lambda)$. [1.0 Ptos.]

Logro 4: Obtener CCR [0.5 Ptos.] y que la distribución es Normal$\left(\lambda,\, \frac{\zeta}{\sqrt{n}}\right)$ [0.5 Ptos.].

Logro 5: Mostrar que EM es consistente. [1.0 Ptos.]

Logro 6: Mostrar que EMV es más eficiente. [1.0 Ptos.]

+ 1 Punto Base

Pregunta 3

Desde el explorador solar del ministerio de energía (https://solar.minenergia.cl/exploracion) se descargo información de la radiación solar que ha afectado al campus San Joaquin UC al mediodía entre 2004 y 2016, y a partir de una muestra aleatoria se construyeron algunos modelos y análisis estadísticos. Entre las variables analizadas están la radiación global (glb) en W/m2, temperatura (temp) a una altura de 2 metros en grados Celsius, la velocidad de viento (vel) en m/s y si existía en ese momento presencia (1: si, 0: no) de nubosidad (cloud).

A continuación se presenta un resumen para la variable `glb`, los coeficientes de determinación R^2 para siete modelos de regresión lineal para predecir `glb` y los valores p de la prueba KS para la normalidad de los residuos de estos modelos:

```
------------------------------------------
     n   mean     sd median   min     max
------------------------------------------
glb 28 714.61 395.78 973.71 47.17 1079.47
------------------------------------------

lm(glb ~ regresores):
--------------------------------------------
modelo       regresores R-squared    p.value
--------------------------------------------
     1             temp    0.5492   0.982598
     2              vel    0.2556   0.745679
     3            cloud    0.8676   0.275897
     4        temp, vel    0.5819   0.981228
     5       cloud, vel    0.8917   0.973311
     6      cloud, temp    0.9249   0.846099
     7 cloud, temp, vel    0.9283   0.642811
--------------------------------------------
```

(a) **[2.0 Ptos.]** Complete la información faltante del modelo 2:

```
lm(formula = glb ~ vel, data = data_aux)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    229.7      175.1   1.312  0.20095
vel            384.4      128.6   X.XXX  0.00606
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: XXXX on 26 degrees of freedom
Multiple R-squared:  0.2556,Adjusted R-squared:  X.XXX
F-statistic: X.XXX on 1 and 26 DF,  p-value: X.XXXXX
```

(b) **[3.0 Ptos.]** Compare el modelo 7 con el mejor modelo simple. ¿El aporte conjunto de las dos variables que se agregan al mejor modelo simple es significativo? Utilice un nivel de significancia del $5\,\%$.

(c) **[1.0 Ptos.]** ¿Cuál de los modelos cumple el supuesto de Normalidad de los residuos de mejor manera? Justifique su respuesta.

Solución

(a) Tenemos que

$$\texttt{t value} = \frac{384.4}{128.6} = 2.989$$

A partir de $n = 28$, $S_Y = 395.78$ y $R^2 = 0.2556$ obtenemos

$$\text{SCT} = (n-1)\cdot S_Y^2 = 4229329 \to \text{SCE} = (1-R^2)\cdot \text{SCT} = 3148312.$$

Luego

$$\texttt{Residual standard error} = S_{Y|X} = \sqrt{\frac{SCE}{n-2}} = 347.9784$$

y

$$\texttt{Adjusted R-squared} = 1 - \frac{S^2_{Y|X}}{S^2_Y} = 0.226969$$

Finalmente

$$\texttt{F-statistic} = \frac{\text{SCR}/1}{\text{SCE}/(n-2)} = \frac{(\text{SCT-SCE})/1}{\text{SCE}/(n-2)} = 8.927458 = \texttt{t value}^2$$

y

$\texttt{p-value} = 0.00606$ (En regresión simple este valor coincide con el `p-value` del `t value` de la pediente)

(b) Por R^2 el mejor modelo de regresión simple, es el modelo 3, cuya SCE es $(1 - 0.8676) \cdot \text{SCT} = 559963.1$. La SCE del modelo 7 es $(1 - 0.9283) \cdot \text{SCT} = 303242.9$.

Del Formulario tenemos que

$$F = \frac{(\text{SCE}_1 - \text{SCE}_2)/r}{\text{SCE}_2/(n-(k+r)-1)} \sim F(r, n-(k+r)-1)$$

Reemplazando $\text{SCE}_1 = 559963.1$, $\text{SCE}_2 = 303242.9$, $n = 28$, $k = 1$ y $r = 2$, se tiene que $F = 10.159$.

Como el valor critico para un 5 % de significancia es $F_{1-0.05}(2, 24) = 3.40$, entonces el aporte conjunto de la variable `vel` y `temp`, en presencia de `cloud`, es significativo.

(c) El mejor ajuste Normal se logra en los residuos del modelo 1, ya que la prueba KS logra el mayor valor-p.

Asignación de puntaje:

Logro 1: `t value` [0.5 Ptos.], `Residual standard error` [0.5 Ptos.]

Logro 2: `Adjusted R-squared` [0.4 Ptos.], `F-statistic` [0.4 Ptos.] y `p-value` [0.3 Ptos.]

Logro 3: Indicar, de manera justificada, que mejor modelo de regresión simple es el modelo 3 [1.0 Ptos.]

Logro 4: Obtener valor $F = 10.159$ [1.0 Ptos.]

Logro 5: Calcular valor critico correctamente [0.5 Ptos.] y concluir que el aporte conjunto es significativo [0.5 Ptos.].

Logro 6: Indicar que el modelo 1 logra los residuos más Normales, debido a que el valor-p del test KS es el mayor entre los calculados. [1.0 Ptos.]

+ 1 Punto Base