Pontificia Universidad Católica de Chile
Facultad de Matemáticas
Departamento de Estadística

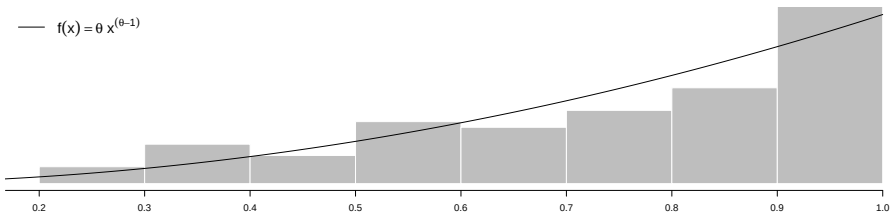
Primer Semestre 2024

Curso : **Probabilidad y Estadística**
Sigla : **EYP1113**
Profesores : **Ana María Araneda L., Ricardo Aravena C., Ricardo Olea O., Felipe Ossa M. e Inés Varas C.**

PAUTA INTERROGACIÓN 4

Problema 1

La estación meteorológica de Quintero, ubicada en la Región de Valparaíso y operada por ENEL Chile S.A., recopila datos meteorológicos vitales para el monitoreo y análisis del clima en la zona. Entre los parámetros registrados se encuentra la humedad relativa del aire, medida cada hora. Los datos históricos de humedad relativa, junto con otros parámetros meteorológicos, como la temperatura ambiente y la velocidad del viento, están disponibles en el portal del Sistema de Información Nacional de Calidad del Aire (SINCA). La siguiente figura muestra el comportamiento de la humedad relativa, medida entre 0 y 1, de una muestra aleatoria de observaciones, y el ajuste de un modelo cuya función de densidad está dada por $f(x) = \theta x^{\theta-1}, 0 \leq x \leq 1$, con $\theta > 0$. Suponga que usted dispone de una muestra aleatoria de n mediciones de humedad relativa del aire.



Asuma que estas mediciones siguen un modelo con la función de densidad propuesta y que son independientes entre sí.

(a) Obtenga el estimador de momentos para el parámetro θ del modelo propuesto.

(b) Obtenga el estimador máximo verosímil (EMV) para el mismo parámetro θ.

(c) Suponga que el verdadero valor de θ es igual a 3. En este caso, ¿cuál es la probabilidad aproximada de que el EMV de θ sea mayor a 2.5, considerando 30 mediciones?

Solución

(a) La esperanza de la humedad relativa en una hora dada corresponde a:

$$\mathsf{E}(X) = \int_0^1 x\ \theta x^{\theta-1}\ dx = \theta \int_0^1 x^{\theta}\ dx = \frac{\theta}{\theta+1}.$$

Igualando 1er momento empírico a teórico se tiene que:

$$\overline{X} = \frac{\theta}{\theta+1} \rightarrow \hat{\theta} = \frac{\overline{X}}{1-\overline{X}}.$$

(b) Para una muestra aleatoria $X_1, \dots X_n$ independiente, la función de verosimilitud para este caso está dada por

$$L(\theta) = \prod_{i=1}^{n} \theta X_i^{\theta-1}.$$

Aplicando logaritmo natural

$$\ln L(\theta) = n\, \ln(\theta) + (\theta - 1) \sum_{i=1}^{n} \ln X_i.$$

Para encontrar el máximo, debemos derivar con respecto a θ y luego igualar a cero para determinar valor crítico.

$$\frac{d \ln L(\theta)}{d\, \theta} = \frac{n}{\theta} + \sum_{i=1}^{n} \ln X_i = 0 \to \theta = -\frac{n}{\sum_{i=1}^{n} \ln x_i}.$$

Notar que

$$\frac{d^2 \ln L(\theta)}{d\, \theta^2} = -\frac{n}{\theta^2} < 0,$$

implica que el valor crítico corresponde a un máximo y por lo tanto, el EMV de θ esta dado por:

$$\widetilde{\theta} = -\frac{n}{\sum_{i=1}^{n} \ln X_i}.$$

(c) Se pide $P\left(\widetilde{\theta} > 2.5\right)$.

Alternativa 1: Tenemos que

$$P\left(\widetilde{\theta} > 2.5\right) = P\left(-\frac{n}{\sum_{i=1}^{n} \ln X_i} > 2.5\right) = P\left(\frac{1}{n}\sum_{i=1}^{n} [-\ln(X_i)] < 0.40\right) = P\left(\overline{Y} < 0.40\right),$$

con

$$Y_i = -\ln(X_i) = g(X_i) \to X_i = e^{-Y_i} = g^{-1}(Y_i).$$

Notemos que $\Theta_{Y_i} = \mathbb{R}^+$, ya que $\Theta_{X_i} = [0, 1]$, y su función de densidad está dada por

$$f_{Y_i}(y) = e{-}y\, \theta\, e^{-(\theta-1)\, y} = \theta\, e^{-\theta\, y} \quad y > 0.$$

Como

$$Y_i \sim \text{Exponencial}(\theta),$$

entonces por Teorema Central del Límite tenemos que

$$\overline{Y} \overset{\text{aprox}}{\sim} \text{Normal}\left(\frac{1}{\theta}, \frac{1}{\theta\sqrt{n}}\right).$$

Reemplaznado $n = 30$ y $\theta = 3$, la probabilidad solicitada es

$$P(\widetilde{\theta} > 2.5) = P(\overline{Y} < 0.4) \approx \Phi\left(\frac{0.4 - 1/3}{0.0609}\right) \approx \Phi(1.09) = 0.8621.$$

Alternativa 2: Tenemos que

$$\widetilde{\theta} \overset{\text{aprox}}{\sim} \text{Normal}\left(\theta, \sqrt{\frac{1}{I(\theta)}}\right),$$

con

$$\mathsf{I}(\theta) = -\mathsf{E}\left[\frac{d^2}{d\,\theta^2}\,\ln L(\theta)\right] = -\mathsf{E}\left(-\frac{n}{\theta^2}\right) = \frac{n}{\theta^2}.$$

Reemplazando $n = 30$ y $\theta = 3$,

$$\widetilde{\theta} \overset{\text{aprox}}{\sim} \text{Normal}\left(3,\ \frac{3}{\sqrt{30}}\right)$$

y por lo tanto la probabilidad solicitada es

$$P(\widetilde{\theta} > 2.5) \approx 1 - \Phi\left(\frac{2.5-3}{3/\sqrt{30}}\right) \approx 1 - \Phi(-0.91) = \Phi(0.91) = 0.8188.$$

Asignación de Puntaje:

Logro 1: Obtener $\mathsf{E}(X) = \dfrac{\theta}{\theta+1}$. **[1.0 Ptos]**

Logro 2: Obtener que el estimador de momentos $\hat{\theta}$ es $\dfrac{\overline{X}}{1-\overline{X}}$. **[1.0 Ptos]**

Logro 3: Por plantear que la log-verosimilitud es

$$\ln L(\theta) = n\,\ln(\theta) + (\theta-1)\sum_{i=1}^{n}\ln X_i.$$ **[1.0 Ptos]**

Logro 4: Obtener que el estimador de máxima verosimilitud $\widetilde{\theta}$ es $-\dfrac{n}{\sum_{i=1}^{n}\ln X_i}$. **[1.0 Ptos]**

Logro 5: Aplicar correctamente el teorema central del límite sobre $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} -[\ln X_i]$ o deducir correctamente la distribución asintótica de $\widetilde{\theta}$. **[1.0 Ptos]**

Logro 6: Por responder que $P(\widetilde{\theta} > 2.5)$ es igual a 0.8621 o 0.8188. **[1.0 Ptos]**

+ 1 Punto Base

Problema 2

En un estudio sobre el salario de los egresados de las carreras de Ingeniería de diferentes instituciones del país se encuestó egresados de la cohorte 2023, registrándose su salario al momento de la encuesta (en UF) y la universidad de egreso (UC / otras). Se enviaron 72 encuestas, de las cuales solo 28 fueron respondidas completamente. A continuación se muestra un resumen de los resultados obtenidos:

		UC	Otras
	n	15	13
Salario UF	mean	44.5	38.5
	sd	8.4	9.2

Realice los tests de hipótesis adecuados para responder cada una de las siguientes preguntas. En cada una de ellas indique: parámetros, hipótesis nula y alternativa, valor del estadístico del test, valor-p, valor crítico y decisión, justificando esta última. Redacte la conclusión en el contexto del problema. En ambos apartados, considere un nivel de significancia $\alpha = 5\,\%$. Si no es posible entregar un valor-p de manera exacta, entregue un intervalo para su valor.

(a) ¿Existe evidencia que permita afirmar que la probabilidad de que un egresado de Ingeniería escogido de manera aleatoria responda a la encuesta es inferior al 50 %?

(b) ¿Es válido afirmar que el salario medio obtenido por los egresados de Ingeniería de la UC es mayor al salario medio obtenido por los egresados de las otras Universidades? Puede asumir que ambos grupos de observaciones provienen de distribuciones Normales, con la misma varianza.

Solución

(a) Sea p la probabilidad de que un egresado de Ingeniería escogido de manera aleatoria responda a la encuesta.

Las hipótesis son:

$$H_0 : p = 0.5 \quad \text{vs} \quad H_a : p < 0.5.$$

La proporción muestral corresponde a

$$\hat{p} = \frac{28}{72} = 0.389.$$

El estadístico del test corresponde a:

$$Z_0 = \frac{0.389 - 0.5}{\sqrt{\dfrac{0.5\,(1 - 0.5)}{72}}} \approx -1.88.$$

El valor crítico corresponde a

$$k_{0.05} = -k_{0.95} = -1.645 > Z_0.$$

El valor-p corresponde a

$$\text{valor-p} = \Phi(-1.88) = 1 - \Phi(1.88) = 1 - 0.9699 = 0.0301.$$

El valor del estadístico del test es menor al punto crítico o, equivalentemente, el valor-p es menor a 5 %, por lo que, con 5 % de significancia, se rechaza H_0, es decir, los datos entregan evidencia para afirmar que la probabilidad de que un egresado de Ingeniería escogido de manera aleatoria responda a la encuesta es inferior a 50 %.

(b) Sean μ_X y μ_Y las medias de los salarios de los ingenieros egresados de la UC y de otras instituciones del país, respectivamente.

Sea σ la desviación estándar común de los salarios, desconocida.

Las hipótesis son:

$$H_0 : \mu_X = \mu_Y \quad \text{vs} \quad H_a : \mu_X > \mu_Y.$$

Se requiere el estimador de σ^2:

$$S_p^2 = \frac{(15-1)8,4^2 + (13-1)9,2^2}{15+13-2} = 77.058,$$

de donde $S_p = 8.78$.

El estadístico del test corresponde a:

$$T_0 = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{44.5 - 38.5}{8.78\sqrt{\frac{1}{15} + \frac{1}{13}}} \approx 1.80.$$

El valor crítico corresponde a $t_{0.95}(15+13-2) = t_{0.95}(26) = 1.706 < T_0$.

Por otra parte,

$$t_{0.975}(26) = 2.056 \rightarrow 0.025 < \text{valor-p} < 0.05.$$

El valor del estadístico del test es mayor al punto crítico o, equivalentemente, el valor-p es menor a 5 %, por lo que, con 5 % de significancia, se rechaza H_0, es decir, los datos entregan evidencia para afirmar que el salario medio obtenido por los egresados de Ingeniería de la UC es mayor al salario medio obtenido por los egresados de las otras Universidades.

Asignación de Puntaje:

Logro 1: Plantear hipótesis nula y alternativa sobre p y determinar el valor del estadístico del test, $Z_0 \approx -1.88$ **[1.0 Ptos]**.

Logro 2: Indicar valor crítico, $k_{0.05} = -1.645$ y calcular valor-p = 0.0301. **[1.0 Ptos]**

Logro 3: Comparar estadístico del test con punto crítico, o comparar valor-p con significancia $\alpha = 5\,\%$, rechazar H_0 y concluir en términos del problema. **[1.0 Ptos]**

Logro 4: Plantear hipótesis sobre μ_X y μ_Y, determinar el valor de $S_p = 8.78$ y el valor del estadístico del test, $T_0 \approx 1.80$. **[1.0 Ptos]**

Logro 5: Indicar valor crítico $t_{0.95}(26) = 1.706 <$ y que $2.5\,\% <$ valor-p $< 5\,\%$. **[1.0 Ptos]**.

Logro 6: Comparar estadístico del test con punto crítico, o comparar valor-p con significancia $\alpha = 5\,\%$, rechazar H_0 y concluir en términos del problema **[1.0 Ptos]**.

+ 1 Punto Base

Problema 3

En el contexto de un análisis del mercado inmobiliario, se busca desarrollar un modelo de regresión simple para entender cómo las características físicas afectan su precio. Se cuenta con los siguientes datos proveniente de una muestra de 47 casas:

```
precio: Precio de la casa en Unidades de Fomento (UF).
mts2: Metros cuadrados construidos.
nhab: Número de habitaciones destinadas a dormitorios.
terreno: Metros cuadrados del terreno.
```

(a) A continuación se presentan dos modelos de regresión simple ajustados en R con algunos valores faltantes que usted debe completar:

```
lm(formula = PRECIO ~ MTS2, data = Data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2154.335    519.262   4.149 0.000146
MTS2        XXXX.XXX      6.034  12.017 1.22e-15

Residual standard error: 812.2 on XXXX degrees of
freedom
```

```
lm(formula = PRECIO ~ NHAB, data = Data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6291.3      791.9   7.945 4.15e-10
NHAB           587.7      230.0   2.555   X.XXXX

Residual standard error: 1557 on XXXX degrees of
freedom
```

Como complemento a la información anterior se presenta un resumen estadístico de la variable `PRECIO`:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.         sd
   4736    6860    8391    8230    9484   10866   1648.116
```

(b) En base a los resultados presentados en (a), ¿cuál de los dos modelos presenta un mejor ajuste? Justifique.

Solución

(a) Modelo 1: `lm(formula = PRECIO ~ MTS2, data = Data)`

Como valor del estadístico del test para $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$ es $T_0 = 12.017$, y su error estándar es igual a 6.034 se tiene que

$$T_0 = \frac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{6.034} = 12.017 \rightarrow \hat{\beta}_1 = 12.017 \cdot 6.034 = 72.51.$$

Por otra parte, dado que son $n = 47$ casas, los grados de libertad del error son $n - 2 = 45$.

Modelo 2: `lm(formula = PRECIO ~ NHAB, data = Data)`

El valor-p corresponde a:

$$\text{valor-p} = 2 \cdot (1 - P(T_{45} \leq |2.555|)) \approx 2 \cdot (1 - \Phi(2.56)) = 2 \cdot (1 - 0.9948) = 0.0104.$$

Nuevamente, los grados de libertad del error son $n - 2 = 45$.

(b) *Alternativa 1*: Frente a las hipótesis $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$, el primer modelo presenta un menor valor-p ($1.22e - 15$ vs 0.0104), por lo que es preferible frente al segundo modelo.

Alternativa 2: La variablidad no explicada, medida a partir del error estándar residual, $s_{Y \mid X}$, es menor en el primer modelo (812.2 vs 1557), por lo que es preferible frente al segundo modelo.

Alternativa 3: El porcentaje de variablidad explicada, medida a partir $r^2 = 1 - \frac{s^2_{Y \mid X}}{s^2_Y}$, es mayor en el primer modelo ($75.7\,\%$ vs $10.8\,\%$), por lo que es preferible frente al segundo modelo.

Asignación de Puntaje:

Logro 1: Encontrar valor de $\hat{\beta}_1 = 72.51$. **[1.5 Ptos]**

Logro 2: Indicar grados de libertad del error $n - 2 = 45$. **[1.5 Ptos]**

Logro 3: Encontrar valor-p igual 0.0104. **[1.5 Ptos]**.

Logro 4: Concluir correctamente al comparar los modelos en términos de valor-p de la pendiente, o el error estándar residual o los coeficiente de determinación r^2. **[1.5 Ptos]**.

+ 1 Punto Base

Problema 4

El índice de compacidad de Gravelius, también conocido como coeficiente de compacidad, es un método utilizado para caracterizar la forma de las cuencas hidrográficas. Este índice se calcula a partir de la relación entre el perímetro de la cuenca y el perímetro de una cuenca teórica circular que tiene la misma área que la cuenca en estudio. Matemáticamente, se define como:

$$K_c = \frac{P}{2\sqrt{\pi A}},$$

donde P es el perímetro de la cuenca y A es el área de la cuenca. Un índice de Gravelius menor a 1.5 indica que la cuenca es de forma ovalada.

Suponga que en cierta zona, el perímetro y área distribuyen Log-Normal con las siguientes características:

	P (en km)	A (en km2)
median	137.50	117.50
c.o.v.	0.45	0.78

Considere que la correlación entre P y A es 0.26, mientras que entre $\ln(P)$ y $\ln(A)$ es 0.30. ¿Cuál es la probabilidad que una cuenca de la zona no sea ovalada?

Solución

Se pide calcular

$$P(K_c > 1.5) = P\left(\ln(K_c) > \ln(1.5)\right).$$

Notemos que

$$\ln(K_c) = \log(P) - 0.5\ln(A) - \ln(2\sqrt{\pi}),$$

donde

$$\ln(P) \sim \text{Normal}(\lambda_1, \zeta_1) \quad \text{y} \quad \ln(A) \sim \text{Normal}(\lambda_2, \zeta_2).$$

Del enunciado se tiene que

$$\lambda_1 = \ln(137.5) = 4.9236, \quad \zeta_1 = \sqrt{\ln(1 + 0.45^2)} = 0.4294$$

y

$$\lambda_2 = \ln(117.5) = 4.7664, \quad \zeta_2 = \sqrt{\ln(1 + 0.78^2)} = 0.6894.$$

Por lo tanto

$$\ln(K_c) \sim \text{Normal}\left(\mu = \lambda_1 - 0.5 \cdot \lambda_2 - \ln(2\sqrt{\pi}),\, \sigma = \sqrt{\zeta_1^2 + 0.5^2 \cdot \zeta_2^2 - 2 \cdot 0.50 \cdot 0.30 \cdot \zeta_1 \cdot \zeta_2}\right).$$

Reemplazando en los valores de los parámetros, $\ln(K_c) \sim \text{Normal}(\mu = 1.2749,\, \sigma = 0.463)$ Luego, la probabilidad pedida corresponde a

$$P(\ln(K_c) > \ln(1.5)) = 1 - \Phi\left(\frac{0.4054651 - 1.2749}{0.463}\right) \approx 1 - \Phi(-1.88) = \Phi(1.88) = 0.9699.$$

Asignación de Puntaje:

Logro 1: Los parámetros de la distribución de $\ln(P)$ **[1.0 Ptos]**.

Logro 2: Los parámetros de la distribución de $\ln(A)$ **[1.0 Ptos]**.

Logro 3: Determinar la media de $\ln(K_c)$ **[1.0 Ptos]**.

Logro 4: Determinar la desviación estándar de $\ln(K_c)$ **[1.0 Ptos]**.

Logro 5: Indicar que la probabilidad solicitada es

$$P(\ln(K_c) > \ln(1.5)) = 1 - \Phi\left(\frac{0.4054651 - 1.2749}{0.463}\right). \quad \textbf{[1.0 Ptos]}$$

Logro 6: Obtener correctamente la probabilidad final 0.9699. **[1.0 Ptos]**.

+ 1 Punto Base