Pontificia Universidad Católica de Chile
Facultad de Matemáticas
Departamento de Estadística

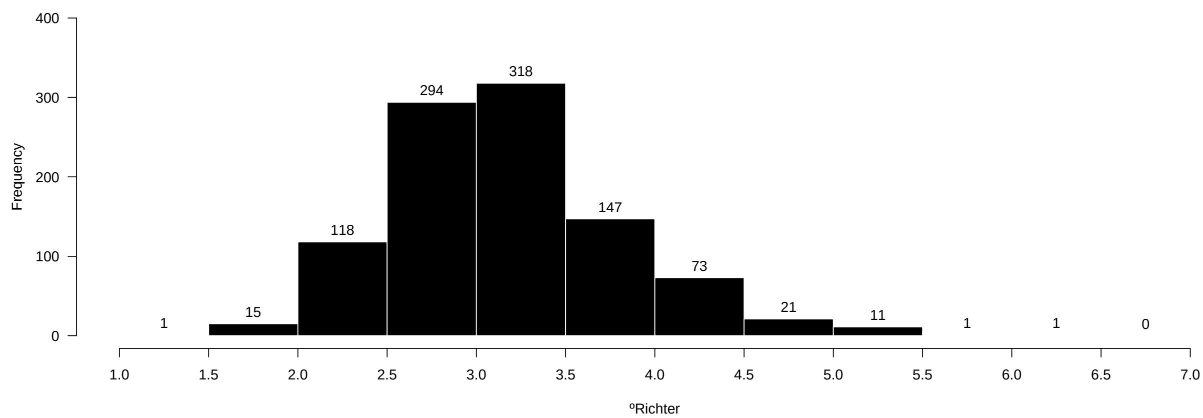
Primer Semestre 2018

Curso	**:**	**Probabilidad y Estadística**
Sigla	**:**	**EYP1113**
Profesores	**:**	**Ricardo Aravena C. y Ricardo Olea O.**

EXAMEN

Problema 1

La siguiente figura muestra el comportamiento empírico de una muestra aleatoria de 1000 sismos ocurridos en Chile entre 2003 y 2015.



Al construir un gráfico de probabilidad Log-Normal y Weibull se obtienen las siguientes rectas

```
           Intercepto Pendiente
Log-Normal   1.147816 0.1999574
Weibull      1.235563 0.1527360
```

(a) **[2.0 Puntos]** Un test de Kolmogorov-Smirnov entrego los siguientes resutados:

```
Kolmogorov-Smirnov test (Log-Normal)
D = 0.055908, p-value = 0.003856
Kolmogorov-Smirnov test (Weibull)
D = 0.11062, p-value = 4.701e-11
```

¿Cuál de los dos modelos ajusta mejor según este test?

(b) **[4.0 Puntos]** Realice ahora un test χ^2 de bondad de ajuste para los siguientes intervalos: 1-2, 2-3, 3-4, 4-5 y 5 $>$. Si fuese necesario colapse intervalos cuando corresponda. ¿Cuál de los dos modelos ajusta mejor según este test? (En sus cálculos redondee al 4to decimal)

Solución

(a) Como el estadístico busca evidencia para rechazar el modelo, el que tiene un valor p mayor es un mejor ajuste, es decir, la Log-Normal ajusta mejor. **[2.0 Ptos]**

(b) A partir del gráfico de probabilidad estimamos los parámetros de los modelos Log-Normal y Weibull

lambda	zeta	eta	beta
1.1478	0.2000	3.4403	6.5472

[0.5 Ptos]

$$H_0 : X \sim \text{Log-Normal} \qquad \text{vs} \qquad H_a : X \not\sim \text{Log-Normal}$$

	Observado	Esperado	(O-E)^/E
1.0-2.0	16	11.6	1.6690
2.0-3.0	412	389.7	1.2761
3.0-4.0	465	481.7	0.5790
4.0-5.0	94	106.6	1.4893
>5.0	13	10.4	0.6500
Total	1000	1000.0	5.6634

[0.5 Ptos]

[0.3 Ptos] Como $X^2 = 5{,}6634 \sim \chi^2(5-1-2) \to 5\,\% < \text{valor-p} < 10\,\%$. **[0.5 Ptos]**

$$H_0 : X \sim \text{Weibull} \qquad \text{vs} \qquad H_a : X \not\sim \text{Weibull}$$

	Observado	Esperado	(O-E)^/E
1.0-2.0	16	28.3	5.3459
2.0-3.0	412	306.7	36.1529
3.0-4.0	465	596.6	29.0288
4.0-5.0	94	68.4	9.5813
>5.0	13	0.0	Inf
Total	1000	1000.0	Inf

[0.5 Ptos]

Como el valor esperado del 5to intervalo es menor a 5, se procede a colapsar los dos últimos

	Observado	Esperado	(O-E)^/E
1.0-2.0	16	28.3	5.3459
2.0-3.0	412	306.7	36.1529
3.0-4.0	465	596.6	29.0288
>4.0	107	68.4	21.7830
Total	1000	1000.0	92.3106

[0.5 Ptos]

[0.3 Ptos] Como $X^2 = 92{,}3106 \sim \chi^2(4-1-2) \rightarrow$ valor-p $< 0{,}5\,\%$. **[0.5 Ptos]**

Por lo tanto, el mejor ajuste se logra con el modelo Log-Normal. **[0.4 Ptos]**

+ 1 Punto Base

Problema 2

Hoy en día, los deportistas de alto rendimiento son analizados estadísticamente con el objetivo de mejorar su desempeño. En específico, en el Mundial que se efectúa en Rusia se podrá ver como monitorean a cada uno de los jugadores, analizando desplazamiento, fuerza motriz y registrando respiración, frecuencia cardiaca, etc.

Usted accede parcialmente a los modelos lineales que ha utilizado la selección argentina para explicar y predecir el rendimiento de sus 24 jugadores (entre los que destaca Messi). En este se busca explicar el desplazamiento km/hrs (Y) durante un encuentro a través de las variables explicatorias: X_1-capacidad respiratoria (en ml/kg), X_2-peso (en kgs), X_3-Talla (en cms) y X_4-potencia (que corresponde a la fuerza explosiva y es medida en mmol/ATP/Kg/seg).

Una primera aproximación se realiza por medio de modelos simples

```
Y ~ X1:
Coefficients    Estimate Std. Error t value
Intercept        X.XXXXX    0.45288   6.817
X1               1.88610    X.XXXXX  25.641

Residual standard error: X.XXXX on XX degrees of freedom
Multiple R-2:  0.9676,    Adjusted R-2:  0.9662
F-statistic: XXX.X on 1 and XX DF
```

```
Y ~ log(X4):
Coefficients    Estimate Std. Error t value
Intercept        1.88195    0.03809   49.41
log(X4)          X.XXXXX    0.01214   XX.XX

Residual standard error: 0.02613 on XX degrees of freedom
Multiple R-2:  X.XXXX,    Adjusted R-2:  X.XXXX
F-statistic: 440.5 on 1 and XX DF
```

(a) **[3.0 Puntos]** Complete la salida **R** de regresión simple. En el caso que algún campo no pueda ser obtenido, indique explícitamente con "falta información". ¿Cuál de los dos modelos es el más apropiado? Justifique.

(b) **[3.0 Puntos]** En una segunda etapa, se ajustan diversos modelos los cuales se presentan en la tabla adjunta:

Modelo	Variables Incluidas	R^2
1	X_1	0,968
2	X_2	0,841
3	X_2, X_3	0,872
4	X_1, X_2, X_3, X_4	0,984

i. ¿Es significativo, al 5 %, el aporte de la talla? Sea explicito, indique hipótesis, test, valor crítico y conclusión.

ii. ¿Es significativo, al 10 %, el aporte conjunto de la capacidad respiratoria y la potencia? Sea explicito, indique hipótesis, test, valor crítico y conclusión.

Respuesta

(a) Por cada valor *[0.2 Ptos]* $\rightarrow$ **[2.4 Ptos]**

```
Y ~ X1:
Coefficients    Estimate Std. Error t value
Intercept        3.08730    0.45288   6.817
X1               1.88610    0.07356  25.641

Residual standard error: 0.3100 on 22 degrees of freedom
Multiple R-2:  0.9676,    Adjusted R-2:  0.9662
F-statistic: 657.46 on 1 and 22 DF
```

```
Y ~ log(X4):
Coefficients    Estimate Std. Error t value
Intercept        1.88195    0.03809   49.41
log(X4)          0.25480    0.01214   20.99

Residual standard error: 0.02613 on 22 degrees of freedom
Multiple R-2:  0.9524,    Adjusted R-2:  0.9501
F-statistic: 440.5 on 1 and 22 DF
```

Mejor ajuste se logra con modelo $Y \sim X_1$, ya que su R^2, r^2, t-value y F-value son mayores. **[0.6 Ptos]**

(b) Notemos que

$$F = \frac{(SCE_{(r)} - SCE)/r}{SCE/(n-k-r-1)} = \frac{(SCE_{(r)} - SCE)/r}{SCE/(n-k-r-1)} \cdot \frac{1/SCT}{1/SCT} = \frac{(R^2 - R^2_{(r)})/r}{(1-R^2)/(n-k-r-1)}$$

i. Se pide comparar modelos 2 vs modelo 3 para responder la siguiente hipótesis **[0.2 Ptos]**

$$\mathrm{H}_0 : \beta_3 = 0 \qquad \text{vs} \qquad \mathrm{H}_a : \beta_3 \neq 0 \quad \textbf{[0.2 Ptos]}$$

El estadístico de prueba bajo H_0 es

$$\textbf{[0.5 Ptos]} \quad F_0 = \frac{(0{,}872 - 0{,}841)/1}{(1-0{,}872)/(24-1-1-1)} = 5{,}086 \sim F(1,21) \quad \textbf{[0.2 Ptos]}$$

Como $F_0 = 5{,}086 > 4{,}32 = F_{0{,}95}(1,21)$ **[0.2 Ptos]**, se rechaza H_0, es decir, el aporte de X_3-Talla es significativo cuando X_2-peso está en el modelo. **[0.2 Ptos]**

ii. Se pide comparar modelos 3 vs modelo 4 para responder la siguiente hipótesis **[0.2 Ptos]**

$$\mathrm{H}_0 : \beta_1 = \beta_4 = 0 \qquad \text{vs} \qquad \mathrm{H}_a : \text{Al menos un } \beta_j \neq 0, \quad j = 1,4 \quad \textbf{[0.2 Ptos]}$$

El estadístico de prueba bajo H_0 es

$$\textbf{[0.5 Ptos]} \quad F_0 = \frac{(0{,}984 - 0{,}872)/2}{(1-0{,}984)/(24-2-2-1)} = 66{,}5 \sim F(2,21) \quad \textbf{[0.2 Ptos]}$$

Como $F_0 = 66{,}5 > 2{,}61 = F_{0{,}90}(2,19)$ **[0.2 Ptos]**, se rechaza H_0, es decir, el aporte conjunto de X_1 y X_5 es significativo cuando X_2 y X_3 están en el modelo. **[0.2 Ptos]**

+ 1 Punto Base

Problema 3

Para este fin de semana se pronostican lluvias y posibles nevazones en gran parte del país, lo que seguramente traerá como consecuencia múltiples cortes de luz. El tiempo que le tomará a los equipos de emergencia de las distintas empresas de distribución eléctrica solucionar los cortes, depende obviamente de la expertise del equipo, cuantificada entre cero y uno. Suponga que la expertise de un equipo cualquiera se comporta como una variable aleatoria Beta$(2, 3)$ y que el tiempo (en horas) condicionado a una expertise x, se comporta como una variable aleatoria Gamma$(3/4,\, x)$. ¿Cuál es el tiempo esperado que le tomaría a un equipo solucionar un corte?

Solución

Definamos como X a la expertise y como Y al tiempo, del enunciado tenemos que

[1.0 Ptos] $\quad X \sim \text{Beta}(2, 3) \quad \text{e} \quad Y \,|\, X = x \sim \text{Gamma}(3/4,\, x)$ **[1.0 Ptos]**

con $\Theta_X = [1,\, 0]$. **[1.0 Ptos]**

Se pide

[1.0 Ptos] $$\text{E}(Y) = \text{E}[\text{E}(Y \,|\, X)] = \text{E}\left(\frac{3}{4\,X}\right) = \int_0^1 \frac{3}{4\,x} \cdot \frac{x\,(1-x)^2}{B(2,3)}\,dx = \frac{3 \cdot 12}{4} \int_0^1 (1-x)^2\,dx = 9 \cdot \frac{1}{3} = 3$$ **[2.0 Ptos]**

+ 1 Punto Base

Problema 4

Para este fin de semana, se espera que la ciudad de Santiago nuevamente sea afectada por un graupel (entre granizo y copo de nieve). A partir del fenomeno ocurrido hace unas semanas se determinó que el comportamiento empíricos de los diámetros de los graupeles se ajustan a un modelo IGAU(θ, β), cuya función de densidad está dada por:

$$f(x) = \frac{1}{x}\,\phi[\ln(x/\theta),\,\beta], \qquad \phi(z,\,\beta) = \frac{\sqrt{\beta}}{\exp(z/2)}\,\phi_{\text{Normal}}\left\{\sqrt{\beta}\left[\frac{\exp(z)+1}{\exp(z/2)}\right]\right\}$$

para $x > 0$, $\theta > 0$ y $\beta > 0$.

El k-ésimo momento de este modelo está dado por $\mathrm{E}(X^k) = \theta^k \sum_{i=0}^{k-1} \frac{(k-1+i)!}{i!\,(k-1-i)!}\left(\frac{1}{2\beta}\right)^i$.

Por simplicidad, suponga que el parámetro β es conocido.

(a) **[3.0 Ptos]** Obtenga el estimador de momento de θ y su distribución aproximada.

(b) **[2.0 Ptos]** Obtenga la distribución aproximada del estimador máximo verosímil.

(c) **[1.0 Ptos]** ¿Son igualmente de eficientes ambos estimadores, según los resultados obtenidos en (a) y (b)?

Respuesta

(a) Como $\mathrm{E}(X) = \theta$, entonces el estimador de momentos está dado por

$$\textbf{[1.0 Ptos]} \quad \hat{\theta} = \overline{X} \overset{\text{aprox}}{\sim} \text{Normal}\left(\theta,\, \sqrt{\frac{\theta^2}{\beta\,n}}\right) \quad \textbf{[2.0 Ptos]}$$

(b) Sean X_1, X_2,..., X_n una muestra aleatoria de diámetros de graupeles cuyo comportamiento será ajustado por una distribución IGAU(θ, β).

La función de verosimilitud y su logarítmo natural están dadas por

$$L(\beta) = \left(\prod_{i=1}^n X_i\right)^{-1} \beta^{n/2} \left(\prod_{i=1}^n X_i\right)^{-1/2} \theta^{n/2}\,(2\,\pi)^{-n/2}\,\exp\left[-\frac{1}{2}\,\frac{\beta}{\theta}\sum_{i=1}^n \frac{(X_i+\theta)^2}{X_i}\right] \quad \textbf{[0.4 Ptos]} \tag{1}$$

$$\ln\,L(\beta) = -\frac{3}{2}\sum_{i=1}^n \ln(X_i) + \frac{n}{2}\,\ln(\beta) + \frac{n}{2}\ln(\theta) - \frac{n}{2}\,\ln(2\,\pi) - \frac{1}{2}\,\frac{\beta}{\theta}\sum_{i=1}^n \frac{(X_i+\theta)^2}{X_i} \quad \textbf{[0.4 Ptos]} \tag{2}$$

Derivando (**??**) parcialmente dos veces con respecto a θ se tiene

$$\textbf{[0.4 Ptos]} \quad \frac{\partial^2}{\partial\theta^2}\,\ln\,L(\beta) = -\frac{n}{2\,\theta^2} - \frac{\beta}{\theta^3}\sum_{i=1}^n X_i \to I(\beta) = -\mathrm{E}\left(-\frac{n}{2\,\theta^2} - \frac{\beta}{\theta^3}\sum_{i=1}^n X_i\right) = \frac{n}{\theta^2}\left(\frac{1}{2}+\beta\right) \quad \textbf{[0.4 Ptos]}$$

Luego, su distribución asintótica está dada por

$$\widetilde{\theta} \overset{\text{aprox}}{\sim} \text{Normal}\left(\theta,\, \sqrt{\frac{\theta^2}{(\beta+1/2)\,n}}\right) \quad \textbf{[0.4 Ptos]}$$

(c) Como ambos estimadores son asintóticamente insesgados, procedemos a comparar sus varianzas: **[0.4 Ptos]**

$$\mathrm{Var}(\widetilde{\theta}) = \frac{\theta^2}{(\beta+1/2)\,n} < \frac{\theta^2}{\beta\,n} = \mathrm{Var}(\hat{\theta}) \quad \textbf{[0.4 Ptos]}$$

Por lo tanto, el EMV es más eficiente para todo valor de β. **[0.2 Ptos]**

+ 1 Punto Base