Pontificia Universidad Católica de Chile
Facultad de Matemáticas
Departamento de Estadística

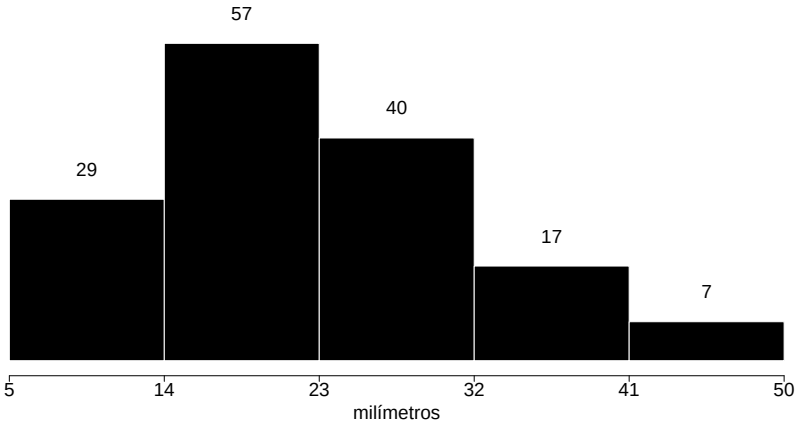
Primer Semestre 2016

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

Pauta Examen

Problema 1

La semifinal que ganó Chile a Colombia durante la Copa América Centenario 2016, fue interrumpida por una tormenta severa. Este tipo de tormentas se caracteriza por durar un corto período de tiempo con fuertes lluvias, fuertes vientos, granizos, rayos y en algunas ocasiones hasta tornados. Una muestra de 150 granizos que cayeron durante el período en que el partido estuvo suspendido, mostró el siguiente comportamiento frecuentista para el diámetro (en milímetros) de estos:



Los primeros tres momentos empíricos fueron: 22.54035, 593.0253 y 17610.05.

¿Estos datos se ajustan a una distribución Weibull(η, $\beta = 2$) traslada en 5 mm.? Estime el parámetro η por el método de momentos (considere dos decimales) y utilice un nivel de significancia del 5 % para su decisión.

Nota: Si $X \sim$ Weibull(η, β) traslada en α, entonces

$$f(x) = \frac{\beta}{\eta}\left[\frac{(x-\alpha)}{\eta}\right]^{\beta-1} \exp\left\{-\left[\frac{(x-\alpha)}{\eta}\right]^{\beta}\right\}, \quad F(x) = 1 - \exp\left[-\left(\frac{x-\alpha}{\eta}\right)^{\beta}\right]$$

con $x \geq \alpha$, $\alpha > 0$, $\eta > 0$ y $\beta > 0$. Para $r > 0$ se tiene que $\mathrm{E}[(X-\alpha)^r] = \eta^r\,\Gamma(1 + r/\beta)$.

Solución

Apliquemos método de momentos para estimar η

[0.5 Ptos.] $\mathrm{E}(X) = \alpha + \eta\,\Gamma\left(1 + \frac{1}{\beta}\right) = 5 + \eta\,\Gamma\left(1 + \frac{1}{2}\right) = 5 + \eta\,\frac{\sqrt{\pi}}{2} = 22.54035 = \overline{X}_n$ **[0.5 Ptos.]**

Despejando

$$\hat{\eta} = 19.79 \qquad \textbf{[0.5 Ptos.]}$$

Se pide contrastar las siguientes hipótesis

$H_0 : X \sim$ Weibull(η, $\beta = 2$) trasladada en 5 mm. vs. $H_a : X \not\sim$ Weibull(η, $\beta = 2$) trasladada en 5 mm. **[0.5 Ptos.]**

Del enunciado, la función de probabilidad acumulada está dada por

$$F_X(x) = 1 - \exp\left[-\left(\frac{x-5}{19.79}\right)^2\right] \qquad \textbf{[0.5 Ptos.]}$$

Luego, las probabilidades de cada intervalo y sus valores observados serían:

	Obs	Prob	Esp	Estadístico
5-14	29	0.18683434	28.025151	0.03390991
14-23	57	0.37592945	56.389418	0.00661136
23-32	40	0.28177886	42.266828	0.12157313
32-41	17	0.11890928	17.836392	0.03922047
41-50	7	0.03654807	5.482211	0.42021074
Total	150	1.00000000	150.000000	0.62152560

[2.0 Ptos.]

El estadístico de prueba X^2 resulto ser igual a 0.62152560 y comparando con el percentil 95 % de una $\chi^2(5-1-1)$ ($c_{0.95}(5-1-1) = 7.81$) **[0.5 Ptos.]**, se concluye que no existe suficiente evidencia para rechazar la hipótesis que el diámetro de los granizos caídos se comportan según una distribución Weibull($\eta, \beta = 2$) trasladada en 5 mm. **[1.0 Ptos.]**

+ 1 Punto Base

Problema 2

Dado que últimamente se ha discutido el efecto "asados" en la contaminación, usted decide llevar a cabo un estudio que permita zanjar científicamente la discusión. Para ello selecciona al azar 45 días de los últimos tres años, de los cuales en 25 se ha observado un alto nivel de contaminación (emergencia) y en los otros 20 los niveles han sido bajos. Además, en estos 45 días hay 12 de ellos en los cuales el día previo la selección ha tenido una confrontación de alto calibre (Clasificatorias, Mundial, Copa América, Copa América Centenario).

Junto a esta información se ha recolectado otras variables relevantes en el modelo de pronóstico (temperatura media, día de la semana, entre otras) y se han ajustado variados modelos.

Si Y corresponde al nivel de contaminante PM 2.5 μm, con las siguientes cuatro variables explicatorias:

X_1: Día previo jugó Chile (1: Si, 0: No)
X_2: Día de la semana (1: Laboral, 0: Festivo)
X_3: Temperatura media prevista (en ^{o}C)
X_4: Nivel de contaminante día previo.

Algunos resultados (obtenidos con R) son los siguintes:

```
            X1     X2     X3    X4     Y
Promedio  0.266  0.622  11.82  67.6  72.2
Desv.Est. 0.447  0.490   4.35  10.1  12.0

lm(formula = Y ~ X1)
Multiple R-squared:  0.1227,    Adjusted R-squared:  0.1023

lm(formula = Y ~ X4)
Multiple R-squared:  0.5945,    Adjusted R-squared:  0.5851

lm(formula = Y ~ X1 + X4)
Multiple R-squared:  0.6615,    Adjusted R-squared:  0.6454

lm(formula = Y ~ X1 + X2 + X3 + X4)
Multiple R-squared:  0.8268,    Adjusted R-squared:  0.8095
```

Lleva a cabo las hipótesis respectivas que permitan responder las siguientes interrogantes en base a los cuatro modelos propuestos.

(a) ¿Es efectivo que si Chile juega el día previo influye en el nivel de contaminante?

(b) ¿Es relevante el aporte conjunto de las variables X_1, X_2 y X_3 al modelo?

Nota: Considere un nivel de significancia del 5 %, sea explícito en el planteamiento de hipótesis, test a utilizar, reglas de decisión y decisión propiamente tal.

Solución

(a) El aporte significativo de X_1 se puede evaluar en:

- Modelo Nulo vs Modelo $Y \sim X_1$.

$$F = \frac{SCR/k}{SCE/(n-k-1)} \sim F(k, n-k-1) \qquad \textbf{[0.2 Ptos.]}$$

donde $k = 1$ **[0.1 Ptos.]** y

$$SCR = R^2 \times SCT = 0.1227 \times (45-1) \cdot 144 = 777.4272 \qquad \textbf{[0.2 Ptos.]}$$
$$SCE = SCT - SCR = (45-1) \cdot 144 - 777.4272 = 5558.573 \qquad \textbf{[0.2 Ptos.]}$$

Reemplazando, se observa que **[0.4 Ptos.]** $F = 6.01402 > 4.07 = F_{0.95}(1, 43)$ **[0.2 Ptos.]**, es decir, el aporte de X_1 es significativo. **[0.2 Ptos.]**

- Modelo $Y \sim X_4$ vs Modelo $Y \sim X_4 + X_1$.

$$F = \frac{(SCE_{(r)} - SCE)/r}{SCE/(n-r-k-1)} \sim F(r, n-r-k-1) \qquad \textbf{[0.2 Ptos.]}$$

donde $k = 1$, $r = 1$ **[0.1 Ptos.]** y

$$SCE_{(r)} = (1 - 0.5945) \times (45-1) \cdot 144 = 2569.248 \qquad \textbf{[0.2 Ptos.]}$$
$$SCE = (1 - 0.6615) \times (45-1) \cdot 144 = 2144.736 \qquad \textbf{[0.2 Ptos.]}$$

Reemplazando, se observa que **[0.4 Ptos.]** $F = 8.313146 > 4.07 = F_{0.95}(1, 42)$ **[0.2 Ptos.]**, es decir, el aporte de X_1 es significativo en presencia de X_4. **[0.2 Ptos.]**

(b) Con la información recibida para cuantificar el aporte se pueden comparar los modelos $Y \sim X_4$ vs $Y \sim X_4 + X_1 + X_2 + X_3$.

$$F = \frac{(SCE_{(r)} - SCE)/r}{SCE/(n-r-k-1)} \sim F(r, n-r-k-1) \qquad \textbf{[0.4 Ptos.]}$$

donde $k = 1$, $r = 3$ **[0.2 Ptos.]** y

$$SCE_{(r)} = (1 - 0.5945) \times (45-1) \cdot 144 = 2569.248 \qquad \textbf{[0.4 Ptos.]}$$
$$SCE = (1 - 0.8268) \times (45-1) \cdot 144 = 1097.395 \qquad \textbf{[0.4 Ptos.]}$$

Reemplazando, se observa que **[0.8 Ptos.]** $F = 17.88299 > 2.84 = F_{0.95}(3, 40)$ **[0.4 Ptos.]**, es decir, el aporte conjunto de X_1, X_2 y X_3 es significativo en presencia de X_4. **[0.4 Ptos.]**

+ 1 Punto Base

Problema 3

En la reciente elección presidencial del Perú, finalmente la diferencie entre los candidatos fue de un 0.32 %. ¿Qué tamaño muestral debería haber tenido una encuesta previa a la elección, para que la estimación del porcentaje de preferencia de un candidato tuviese ese margen de error? Utilice una confianza del 95 % y el criterio de varianza máxima.

Solución

Consideremos X_1, X_2, ..., X_n una muestra aleatoria Bernoulli(p), donde p representan el porcentaje de preferencia por un candidato cualquiera.

Para el caso de varianza máxima, el error de estimación de un intervalo de confianza al 95 % está dado por

[4.0 Ptos.] $$\omega = \frac{1.96}{2\sqrt{n}} = 0.0032$$ **[1.0 Ptos.]**

Despejando se tiene que

$$n \approx 93789.06$$ **[1.0 Ptos.]**

Por lo tanto, una encuesta de aproximadamente noventa y tres mil setecientos noventa personas sería capaz de estimar con un margen de error igual a la diferencia real observada.

+ 1 Punto Base

Problema 4

En las recientes primarias se vio una menguada concurrencia. Para matar el aburrimiento dos alumnos del curso, que fueron designados vocales de las comunas de Ñuñoa (Nueva Mayoría) y Macul (ChileVamos), apostaron respecto a dos atributos: participación femenina y tiempo que los votantes demoran en el proceso.

Para efectos de proteger la identidad de los partidarios designaremos a cada comuna como A y B.

Característica	Comuna A	Comuna B
Número de votantes	48	34
Número de Mujeres	26	12
Tiempo promedio (seg)	85	92

(a) ¿Existe evidencia que permita afirmar una mayor participación femenina en la comuna A en relación a la comuna B? Considere un nivel de significancia del 5 %.

(b) Asumiendo que los tiempos se comportan de acuerdo a una distribución Exponencial, ¿se puede afirmar que los tiempos medios que los votantes demoran en el proceso difieren entre la comuna A y comuna B? Considere un nivel de significancia del 5 %.

Solución

(a) Se pide contrastar las siguientes hipótesis

$$\mathrm{H}_0 : p_A = p_B \quad \text{vs.} \quad \mathrm{H}_a : p_A > p_B \qquad \textbf{[0.5 Ptos.]}$$

De la tabla se tiene que

$$\textbf{[0.5 Ptos.]} \qquad \hat{p}_A = \frac{26}{48} = 0.5417 \quad \text{y} \quad \hat{p}_B = \frac{12}{34} = 0.3529 \qquad \textbf{[0.5 Ptos.]}$$

Bajo el supuesto que H_0 es correcta,

$$Z_0 = \frac{(\hat{p}_A - \hat{p}_B)}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{48}+\frac{1}{34}}} \overset{\text{aprox.}}{\sim} \text{Normal}(0,1) \qquad \textbf{[0.5 Ptos.]}$$

donde $\hat{p} = \dfrac{26+12}{48+34} = 0.4634.$ **[0.5 Ptos.]**

Reemplazando, Z_0 es igual a 1.689 que es mayor al percentil 95 % ($k_{0.95} = 1.645$). Por lo tanto, existe evidencia suficiente para rechazar H_0, es decir, en la comuna A hubo una mayor participación femenina en relación a la comuna B. **[0.5 Ptos.]**

Alternativamente, se deduce que valor-p $= 1 - \Phi(1.689) = 0.04 < 0.05$. Por lo tanto, existe evidencia suficiente para rechazar H_0, es decir, en la comuna A hubo una mayor participación femenina en relación a la comuna B. **[0.5 Ptos.]**

(b) Se pide contrastar las siguientes hipótesis

$$\mathrm{H}_0 : \mu_A = \mu_B \quad \text{vs.} \quad \mathrm{H}_a : \mu_A \neq \mu_B \qquad \textbf{[0.5 Ptos.]}$$

Bajo el supuesto que H_0 es correcta,

$$Z_0 = \frac{(\overline{X}_A - \overline{X}_B)}{\hat{\mu}\sqrt{\frac{1}{48}+\frac{1}{34}}} \overset{\text{aprox.}}{\sim} \text{Normal}(0,1) \qquad \textbf{[0.5 Ptos.]}$$

donde $\hat{\mu} = \dfrac{48 \times 85 + 34 \times 92}{48 + 34} = 87.9$. **[0.5 Ptos.]**

Reemplazando, Z_0 es igual a -0.3552632 **[0.5 Ptos.]** que es mayor al percentil 2.5 % ($k_{0.025} = -1.96$) **[0.5 Ptos.]**. Por lo tanto, no existe evidencia suficiente para rechazar H_0, es decir, se puede afirmar que los tiempos medios que los votantes demoran en el proceso difieren entre la comuna A y comuna B. **[0.5 Ptos.]**

Alternativamente, se deduce que valor-p $= 2\,[1 - \Phi(|-0.3552632|)] = 0.7188471 > 0.05$ **[1.0 Ptos.]**. Por lo tanto, no existe evidencia suficiente para rechazar H_0, es decir, se puede afirmar que los tiempos medios que los votantes demoran en el proceso difieren entre la comuna A y comuna B. **[0.5 Ptos.]**

+ 1 Punto Base