Analizar #4 MA340-0 Prim'08 [[RA; HO, VR]]

Intervalos de confianza
Test. de Hipótesis

Problema 1 En una aseguradora, el número de partes recibidos por siniestros en un día sigue una distribución de Poisson. Se desea contrastar las hipótesis

$H_0: \lambda = 2$ vs $H_1: \lambda = 1$

Analizar, usando el lema de Neyman-Pearson, si existe la mejor región crítica con un nivel de significancia del 3%. Para efectuar el contraste, se ha extraído una MAS de 100 días, siendo la media muestral = 1,3.

Sol La receta para aplicar Neyman-Pearson es

- Imponer la razón de verosimilitudes mayor a cierto k
- Llegar a cierto estadístico $T \geq$ ó $\leq$ a cierto k'.
- El k' se obtiene de $\mathbb{P}(T \geq k' / H_0) = \alpha$, o $\mathbb{P}(T \leq k' / H_0) = \alpha$ (según el punto anterior)

En este caso, la razón de verosimilitud para $(x_1, \dots, x_n)$ será

$$r = \frac{f_n(x_1, \dots, x_n / H_1)}{f_n(x_1, \dots, x_n / H_0)} = \frac{e^{-\lambda_1 n} (\lambda_1)^{\sum_{i=1}^n x_i} / \prod_{i=1}^n x_i!}{e^{-\lambda_0 n} (\lambda_0)^{\sum_{i=1}^n x_i} / \prod_{i=1}^n x_i!}$$

$$= e^{-n(\lambda_1 - \lambda_0)} \left(\frac{\lambda_1}{\lambda_0}\right)^{\sum_{i=1}^n x_i} \geq k \quad (*)$$

Imponer

El estadístico será $T(x_1, \dots, x_n) = \sum_{i=1}^n x_i$. Entonces, aplicando log a $(*)$, los $\vec{x} = (x_1, \dots, x_n) \in \mathcal{X}$ que satisfacen $(*)$ satisfarán

$$-n(\lambda_1 - \lambda_0) + \left(\sum_{i=1}^n x_i\right) \ln(\lambda_1/\lambda_0) \geq \ln(k)$$

En este caso, $\lambda_1 = 1$, $\lambda_0 = 2 \Rightarrow \ln(\lambda_1/\lambda_0) = \ln(1/2) < 0$

$$\Rightarrow \sum_{i=1}^n x_i \leq \frac{\ln(k) + n(\lambda_1 - \lambda_0)}{\ln(\lambda_1/\lambda_0)} = k'$$

Entonces, la Región crítica será

$$R^* = \left\{ (x_1, \dots, x_n) \in \mathbb{R}_+^n \;/\; \sum_{i=1}^n x_i \leq k' \right\}$$

¿De donde obtenemos k'? de suponer $\mathbb{P}(R^* / H_0) = \alpha$.

$$\mathbb{P}\left(\sum_{i=1}^n x_i \leq k' / \lambda = 2 \right) = \alpha = 0{,}03$$

¿Como se encuentra k'? Se necesita la distribución de $\sum_{i=1}^n x_i$. Como cada $X_i \sim \text{Poisson}(\lambda)$, $\sum_{i=1}^n X_i \sim \text{Poisson}(n\lambda)$ (si son indep.)

Ahora, bajo H_0, $\lambda = 2 \Rightarrow \sum_{i=1}^n X_i \sim \text{Poisson}(2n)$

como en el ejemplo $n = 100 \Rightarrow \sum_{i=1}^n X_i \sim \text{Poisson}(200)$

Entonces, k' va a ser el primer entero tal que

" $\mathbb{P}(\text{Poisson}(200) \leq k') \geq 0{,}03$ "

Calculado por computador, $k' = 174$ $(\mathbb{P}() = 0{,}0335)$

si $k' = 173$, $\mathbb{P}() = 0{,}0263$

Comparando, como $\overline{X}_n = 1{,}3 = \frac{\sum X_i}{n} \Rightarrow \sum_{i=1}^n X_i = 1300$

se tiene que $\sum_{i=1}^n X_i = 1300 > 174$, por lo que no se rechaza $\lambda = 2$

Problema 2 Sea $X_1, \ldots, X_n$ una MAS del modelo uniforme sobre el intervalo $(0,\beta)$. Escribir un intervalo de confianza para $\beta \in (0,\infty)$, si se quiere un nivel de confianza $1-\alpha$. En particular, escribir el intervalo para $\alpha = 0{,}1$ si se obtiene una muestra de 4 elementos: $1{,}32$, $1{,}13$; $0{,}67$; $0{,}27$. Ind: use como fn. pivote $T(\vec{X},\beta) = \max\{X_1, \ldots, X_n\}/\beta$.

Sol Se necesita un intervalo (a,b) tal que $\mathbb{P}(a \le \beta \le b) = 1-\alpha$.
Para ello usamos la función pivote T (cuya distribución no depende del parámetro β). Calculamos la distribución de T (Def $Y = \max\{X_1, \ldots, X_n\}$).

$$\mathbb{P}(T \le t) = \mathbb{P}(Y/\beta \le t) = \mathbb{P}(Y \le t\beta) \qquad t\beta \in (0,\beta) \Rightarrow t \in (0,1)$$

$$= \mathbb{P}(\max\{X_1, \ldots, X_n\} \le t\beta)$$

$$= \mathbb{P}(X_1 \le t\beta, \ldots, X_n \le t\beta) = \mathbb{P}(X_1 \le t\beta) \cdots \mathbb{P}(X_n \le t\beta)$$

$$= \mathbb{P}(X_1 \le t\beta)^n = \left(\frac{t\beta}{\beta}\right)^n = t^n \qquad t \in (0,1)$$

Entonces, usando la fn pivote, buscamos $(\tilde{a}, \tilde{b})$ tal que para luego encajonar β.

$$\mathbb{P}(\tilde{a} \le T \le \tilde{b}) = 1-\alpha$$

$$\Rightarrow \begin{cases} \mathbb{P}(T < \tilde{a}) = \alpha/2 \\ \mathbb{P}(T > \tilde{b}) = \alpha/2 \end{cases}$$ así los elegimos! (podría elegirse $\tilde{b} = 1$, $\tilde{a}$ tq $\mathbb{P}(T \le \tilde{a}) = \alpha$)

Entonces, $\mathbb{P}(T < \tilde{a}) = \tilde{a}^n = \alpha/2 \Rightarrow \boxed{\tilde{a} = \sqrt[n]{\alpha/2}}$

$$\mathbb{P}(T > \tilde{b}) = 1 - \mathbb{P}(T \le \tilde{b}) = 1 - \tilde{b}^n = \alpha/2 \Rightarrow \boxed{\tilde{b} = \sqrt[n]{1-\alpha/2}}$$

Ahora, encajonamos a β; usamos la definición de $T = Y/\beta$

$$\left.\begin{aligned} &\tilde{a} \le T \Leftrightarrow \tilde{a}\beta \le Y \Leftrightarrow \beta \le Y/\tilde{a} \\ &T \le \tilde{b} \Leftrightarrow Y \le \beta\tilde{b} \Leftrightarrow Y/\tilde{b} \le \beta \end{aligned}\right\} \Rightarrow \text{con Prob } 1-\alpha, \quad \frac{Y}{\tilde{b}} \le \beta \le \frac{Y}{\tilde{a}}$$

Entonces, el intervalo (a,b) pedido será: $a = \frac{Y}{\tilde{b}}$, $b = \frac{Y}{\tilde{a}}$

En el caso específico de este problema, $Y = \max\{x_1, \dots, x_n\} = 1{,}32$; $\alpha = 0{,}1$, $n = 4$. Entonces,

$$\tilde{a} = \sqrt[4]{0{,}05} = 0{,}47$$
$$\tilde{b} = \sqrt[4]{1-0{,}05} = 0{,}99$$

$\Rightarrow a = 1{,}34$, $b = 2{,}79$

Problema 3 Sean $X = (X_1, \dots, X_n)$, $Y = (Y_1, \dots, Y_m)$ muestras aleatorias mutuamente independientes provenientes de las distribuciones $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$ respect., donde σ^2 es desconocida.

a) Proponga un test para contrastar $H_0: \mu_1 \geq \mu_2$ vs $H_1: \mu_1 < \mu_2$

b) ¿Cual es su decisión con un error de tipo I del 5%? Considerando los datos $n = 50$, $m = 60$, $\overline{X}_n = 3{,}2$, $\overline{Y}_m = 3{,}7$, $\frac{1}{n}\sum(x_i - \overline{x}_n)^2 = 0{,}62$, $\frac{1}{m}\sum(y_i - \overline{y}_m)^2 = 0{,}79$.

c) ¿Como cambia la región crítica cuando n y m decrecen?

a/ Notar que las hipotesis se pueden escribir como
$$H_0: \mu_1 - \mu_2 \geq 0 \quad ; \quad H_1: \mu_1 - \mu_2 < 0$$

Obs: $\overline{X}_n \sim N(\mu_1, \frac{\sigma^2}{n})$, $\overline{Y}_m \sim N(\mu_2, \frac{\sigma^2}{m})$

Definiendo $W = \overline{X}_n - \overline{Y}_m$, $W \sim N(\mu_1 - \mu_2, \frac{\sigma^2}{n} + \frac{\sigma^2}{m})$

Si $\overline{X}_n - \overline{Y}_m$ es < 0 como $\overline{X}_n$ estima a μ_1 e $\overline{Y}_n$ estima a μ_2, debiera rechazarse H_0. (esta es la idea de la construcción del test). Por ello, la observación 2).

Como desconocemos σ, hay que ocupar la t-student

$$Z = \frac{\overline{X}_n - \overline{Y}_m - (\mu_1 - \mu_2)}{\sqrt{\sigma^2/n + \sigma^2/m}} \sim N(0,1)$$

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

$$S_Y^2 = \frac{1}{m}\sum_{j=1}^{m}(Y_j - \overline{Y}_m)^2$$

y $S^2 = \dfrac{n S_x^2 + m S_y^2}{n+m}$ es estimador de σ^2

$$\Rightarrow Q = \frac{(n+m)S^2}{\sigma^2} \sim \chi^2_{n+m-2}$$

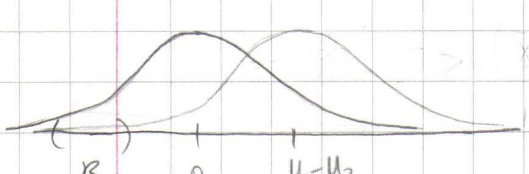
Entonces $\dfrac{Z}{\sqrt{Q/(n+m-2)}} \sim t_{n+m-2}$

$$T = \frac{Z}{\sqrt{Q/n+m-2}} = \frac{\overline{X}_n - \overline{Y}_m - (\mu_1 - \mu_2)}{\cancel{\sigma}\sqrt{\frac{1}{n}+\frac{1}{m}}} \cdot \frac{\sqrt{n+m-2}}{\sqrt{n S_x^2 + m S_y^2}/\cancel{\sigma}}$$

$$= \frac{\left[\overline{X}_n - \overline{Y}_m - (\mu_1 - \mu_2)\right]\sqrt{n+m-2}}{\sqrt{n S_x^2 + m S_y^2}\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t_{n+m-2}$$

Ahora, dos observaciones:

1) Consideramos el caso extremo para la hipótesis nula $\mu_1 - \mu_2 = 0$ ¿Porque?
Al considerar la familia de distribuciones normales de media $\mu_1 - \mu_2 \geq 0$, la región de rechazo debiera estar "hacia el otro lado". Entonces (viendo el dibujo), se tiene que la distribución con $\mu_1 - \mu_2 = 0$ asigna más probabilidad a R (región de rechazo), que es justamente lo que se busca: poner una cota (máxima) para el error tipo I.



2) Consideramos una región de la forma $R = \{(\vec{x}, \vec{y}) \in \mathbb{R}^{n \times m} \,/\, \overline{X}_n - \overline{Y}_m \leq k\}$

Ahora, $\mathbb{P}(R / H_0^*) = \alpha$, con $H_0^*: \mu_1 - \mu_2 = 0$

Entonces, $\mathbb{P}\left(\bar{X}_n - \bar{Y}_m \leq k\right) = \mathbb{P}\left(T \leq \underbrace{\frac{k\sqrt{n+m-2}}{\sqrt{nS_x^2+mS_y^2}\sqrt{\frac{1}{n}+\frac{1}{m}}}}_{\tilde{k}}\right) = \alpha$

con $T \sim t_{n+m-2}$.

En el ejemplo, $T \sim t_{108}$. Así, $\tilde{k} \cong -1{,}645$

$$k = \tilde{k} \cdot \frac{\sqrt{nS_x^2+mS_y^2}}{\sqrt{n+m-2}}\sqrt{\frac{1}{n}+\frac{1}{m}} = -1{,}645 \cdot \frac{\sqrt{50 \cdot 0{,}62 + 60 \cdot 0{,}79}}{\sqrt{108}} \cdot \sqrt{\frac{1}{50}+\frac{1}{60}}$$

$$= -0{,}268$$

Entonces, $\bar{X}_n - \bar{Y}_m = -0{,}5 < -0{,}268$, por lo que se rechaza H_0.

Problema 4 Sea $X=(X_1,\dots,X_n)$ m.a.s. proveniente de la distribución Beta de parámetros $(\theta,1)$, con función de densidad de probabilidad dada por $f(x|\theta)=\theta x^{\theta-1}$; $x\in(0,1)$, donde $\theta>0$.

(a) Muestre que el test más potente de tamaño α para contrastar $H_0:\theta=\theta_0$ vs $\theta=\theta_1$, con $\theta_1>\theta_0$ rechaza H_0 para valores grandes del estadístico $T(X)=T=\sum_{i=1}^{n}\ln X_i$, i.e., tiene una región crítica de la forma

$$R^*=\{x\in[0,1]^n : T(x)\geq k_\alpha\}$$

(b) Dem. que $\underset{\text{bajo } H_0}{-2\theta\ln(X_i)}\sim\chi^2_2$. Deduzca la dist de $-2\theta\sum_{i=1}^{n}\ln X_i$.

(c) ¿Cuál es la decisión en relación a H_0? considerando $\theta_0=2$, $\theta_1=2{,}8$, $\alpha=0{,}05$, $n=10$, $\sum\log x_i=-5{,}846$ y un error de tipo I de 5%. ¿Cuál es el error de tipo II? ¿El p-valor del test es mayor o menor que 5%?

sol a) Nuevamente, Neyman-Pearson.

$$r=\frac{f_n(x_1,\dots,x_n/\theta=\theta_1)}{f_n(x_1,\dots,x_n/\theta=\theta_0)}=\frac{\theta_1^n\left(\prod_{i=1}^{n}x_i\right)^{\theta_1-1}}{\theta_0^n\left(\prod_{i=1}^{n}x_i\right)^{\theta_0-1}}=\left(\frac{\theta_1}{\theta_0}\right)^n\left(\prod_{i=1}^{n}x_i\right)^{\theta_1-\theta_0}\geq k$$

Aplicando $\ln$, $n\ln(\theta_1/\theta_0)+(\theta_1-\theta_0)\ln\left(\prod_{i=1}^{n}x_i\right)\geq\ln(k)$ $\quad\theta_1-\theta_0>0$

$$\ln\left(\prod_{i=1}^{n}x_i\right)=\sum_{i=1}^{n}\ln x_i\geq\frac{\ln(k)-n\ln(\theta_1/\theta_0)}{\theta_1-\theta_0}=k_\alpha$$

Esto define R^*

b) Según el TCV, si $r:(0,1)\to\mathbb{R}_+$, $x\mapsto r(x)=-2\theta\ln(x)$, entonces $s:\mathbb{R}_+\to(0,1)$, $y\mapsto s(y)=e^{-\frac{y}{2\theta}}$ es la inversa de r, y $\frac{\partial s}{\partial y}(y)=e^{-y/2\theta}\cdot\frac{-1}{2\theta}\Rightarrow\left|\frac{\partial s}{\partial y}(y)\right|=\frac{e^{-y/2\theta}}{2\theta}$

Entonces, $f_Y(y)=f_X(s(y))\left|\frac{\partial s}{\partial y}(y)\right|=\cancel{\theta}\left(e^{-y/2\theta}\right)^{\theta-1}\frac{e^{-y/2\theta}}{2\cancel{\theta}}=\frac{e^{-\frac{y}{2\theta}\theta}e^{+\frac{y}{2\theta}}e^{-y/2\theta}}{2}$

$$=\frac{1}{2}e^{-y/2}$$

La densidad de una χ^2_n es

$$g_n(y) = \frac{y^{n/2-1} e^{-y/2}}{\Gamma(n/2) 2^{n/2}}, \text{ si } n=2, \text{ se tiene}$$

$$g_2(y) = \frac{y^{2/2-1} e^{-y/2}}{\Gamma(2/2) 2^{2/2}} = \frac{e^{-y/2}}{2}$$

Como las $(X_i)_{i=1}^n$ son indep, $(-2\theta \ln(X_i) = Y_i)_{i=1}^n$ son indep. Así, la suma de n χ^2_β da $\chi^2_{n\beta}$.

$$\Rightarrow -2\theta \sum_{i=1}^n \ln X_i \sim \chi^2_{2n}$$

c) Para encontrar K_α, necesitamos que $\mathbb{P}(\underbrace{T \geq K_\alpha}_{R^*}) = \alpha$, con $\sum_{i=1}^n \ln X_i = T$

Reescribiendo, $-2\theta \sum_{i=1}^n \ln X_i = -2\theta T \sim \chi^2_{2n}$

y en R^*, $T \geq K_\alpha \Leftrightarrow -2\theta T \leq -2\theta K_\alpha = \tilde{K}$

Entonces, $\mathbb{P}(T \geq K_\alpha / H_0) = \mathbb{P}(\chi^2_{2n} \leq \tilde{K} / H_0) = \alpha$

si $\alpha = 0{,}05$, $n=10$, $\sum_{i=1}^n \ln x_i = -5{,}846$, $\theta_0 = 2$, $\theta_1 = 2{,}8$

$\mathbb{P}(\chi^2_{20} \leq \tilde{K}) = 0{,}05 \Leftrightarrow \tilde{K} = 10{,}85 \Rightarrow K_\alpha = -\frac{\tilde{K}}{2\theta} = -2{,}71$ $\quad H_0$ $\theta = 2$

Así, como $\sum_{i=1}^n \ln X_i = -5{,}846 \leq K_\alpha$, no se rechaza H_0.