Examen de Estadística II. 08-07-2009.

1. X, tiempo de espera en minutos.
$X \sim Ex(\theta),\ f(x) = \theta \exp(-\theta x),\ x > 0,\ \theta > 0.$
$(x_1, \ldots, x_n)$ m.a.s. tal que $\overline{x} = 14,\ s = 12.$

a) La mejor estimación para la media es la proporcionada por el estimador máximo verosímil que se obtiene maximizando la correspondiente función de verosimilitud Exponencial,

$$l(\theta) = \prod_{i=1}^{n} f(x_i; \theta) = \prod_{i=1}^{n} \theta \exp(-\theta x_i) = \theta^n \exp\left(-\theta \sum_{i=1}^{n} x_i\right).$$

De modo que la log-verosimilitud es $\log l(\theta) = n \log \theta - \theta \left(\sum_{i=1}^n x_i\right)$ y su primera derivada resulta $\frac{d \log l(\theta)}{d\theta} = \frac{n}{\theta} - \sum_{i=1}^n x_i$, que igualada a cero proporciona la solución $\hat{\theta} = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\overline{x}}$. Efectivamente, el inverso de la media muestral es el máximo de $\log l(\theta)$ (y por la inyectividad de la función logarítmica, también de $l(\theta)$), pues hace negativa su segunda derivada, $\frac{d^2 \log l(\theta)}{d\theta^2} = -\frac{n}{\theta^2} < 0,\ \forall \theta > 0$, luego la media muestral $\overline{X}$ es el estimador máximo verosímil de la media $\mu = \frac{1}{\theta}$ de una población Exponencial y, en este caso, $\overline{x} = \underline{14}$ es la estimación máximo verosímil.

b) Un estimador es eficiente cuando es insesgado y su varianza alcanza la cota de Cramer-Rao. La media muestral resulta siempre un estimador insesgado de la media poblacional pues $E\left(\overline{X}\right) = E\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{1}{n}\sum_{i=1}^n E(X_i) = \frac{1}{n}\sum_{i=1}^n \mu = \frac{n\mu}{n} = \mu,\ \forall \mu \in \mathbb{R}.$

Para demostrar la condición acerca de la varianza hemos de comprobar que $Var\left(\overline{X}\right) = \frac{1}{I(\mu)},\ \forall \mu > 0$, donde $I(\mu) = E\left[\left(\frac{d \log f(X_1,\ldots,X_n;\mu)}{d\mu}\right)^2\right] = -E\left[\frac{d^2 \log f(X_1,\ldots,X_n;\mu)}{d\mu^2}\right]$.

La varianza de la media muestral es siempre igual a la varianza poblacional dividido entre el tamaño muestral. En el caso de esta m.a.s. Exponencial, $Var\left(\overline{X}\right) = Var\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^n Var(X_i) = \frac{1}{n^2}\sum_{i=1}^n \frac{1}{\theta^2} = \frac{n}{n^2\theta^2} = \frac{1}{n\theta^2} = \frac{\mu^2}{n}$.

Por su parte, la función de densidad conjunta Exponencial es igual a la función de verosimilitud obtenida en el apartado anterior, $f(X_1, \ldots, X_n; \mu) = \mu^{-n} \exp\left(-\mu^{-1}\sum_{i=1}^n X_i\right)$. La log-verosimilitud es $\log f(X_1, \ldots, X_n; \mu) = -n \log \mu - \frac{1}{\mu}\sum_{i=1}^n X_i$ y su primera derivada $\frac{d \log f(X_1,\ldots,X_n;\mu)}{d\mu} = -\frac{n}{\mu} + \frac{\sum_{i=1}^n X_i}{\mu^2} = \frac{n}{\mu^2}\left(\overline{X} - \mu\right)$. De este modo la información de Fisher resulta $I(\mu) = E\left[\left(\frac{d \log f(X_1,\ldots,X_n;\theta)}{d\theta}\right)^2\right] = \frac{n^2}{\mu^4}E\left[(\overline{X} - \mu)^2\right] = \frac{n^2}{\mu^4}Var(\overline{X}) = \frac{n^2}{\mu^4}\frac{\mu^2}{n} = \frac{n}{\mu^2}$ y efectivamente la cota de Cramer-Rao resulta igual a la varianza de $\overline{X}$ c.q.d.

2. X, consumo de gas-rei cada 100 km., en litros.
Y, consumo de gas-futurate cada 100 km., en litros.
$X \sim N(\mu_x, 0.66), (x_1, \ldots, x_{15})$ m.a.s. tal que $\overline{x} = 13.24$.
$(y_1, \ldots, y_{85})$ m.a.s. tal que $\overline{y} = 13.12, s_y^2 = 0.65$.

Asumiendo independencia entre X e Y, esto es, que nada tiene que ver el consumo de gas-rei con el de gas-futurate y que el tamaño muestral $n_y = 85$ es suficientemente grande, se tiene que $\frac{\overline{X}-\overline{Y}-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{s_y^2}{n_y}}} \dot{\sim} N(0,1)$. Por tanto, el intervalo de confianza $1-\alpha$ para $\mu_x-\mu_y$ es $\left[\overline{x}-\overline{y}-z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_x^2}{n_x}+\frac{s_y^2}{n_y}}, \overline{x}-\overline{y}+z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_x^2}{n_x}+\frac{s_y^2}{n_y}}\right]$, donde $z_{\frac{\alpha}{2}}$ es el cuantil de la distribución $N(0,1)$ que deja a su derecha una probabilidad igual a $\frac{\alpha}{2}$.

Por consiguiente $\left[13.24-13.12-1.96\sqrt{\frac{0.66}{15}+\frac{0.65}{85}}, 13.24-13.12+1.96\sqrt{\frac{0.66}{15}+\frac{0.65}{85}}\right] = \underline{[-0.3254, 0.5654]}$ es el intervalo de confianza 0.95 para $\mu_x-\mu_y$, habiendo sustituido $z_{0.025} = 1.96$.

3. X, gasto de los valencianos en las rebajas, en euros.
$(x_1,\ldots,x_{200})$ m.a.s. tal que $\overline{x}=110$, $s=40$.

Denotando por μ a la media poblacional se ha de resolver el siguiente contraste de hipótesis

$$\begin{aligned} H_0 &: \mu \le \mu_0 = (0.85)(125) = 106.25 \\ H_1 &: \mu > \mu_0 = 106.25. \end{aligned}$$

Como el tamaño muestral $n=200$ es elevado y $(x_1,\ldots,x_{200})$ es una m.a.s. se tiene que $\sqrt{n}\,\frac{\overline{X}-\mu}{S} \dot{\sim} N(0,1)$, de modo que el test con un nivel de significación α que se propone para resolver el contraste planteado es

$$\text{rechazar } H_0 \Leftrightarrow \overline{x} \ge \mu_0 + z_\alpha\,\frac{s}{\sqrt{n}}.$$

El nivel de significación crítico se obtiene resolviendo para α la ecuación asociada a la inecuación anterior, esto es, $110 = 106.25 + z_\alpha\,\frac{40}{\sqrt{200}} \Leftrightarrow z_\alpha = 1.3258 \Leftrightarrow \underline{\alpha_0 = 0.0925}$.

4. X_1, efecto del fertilizante F_1.
X_2, efecto del fertilizante F_2.

a) $H_0:$ X_1 e X_2 homogéneas
$H_1:$ no H_0.
Bajo ciertas condiciones genéricas y para tamaños muestrales grandes se tiene que

$$Q = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - n_i\,\frac{n_{\cdot j}}{n}\right)^2}{n_i\,\frac{n_{\cdot j}}{n}} \dot{\sim} \chi^2_{(r-1)(s-1)}.$$

El test que resuelve el contraste de homogeneidad es rechazar $H_0 \Leftrightarrow q \ge \chi^2_{(r-1)(s-1),\alpha}$, donde $\chi^2_{(r-1)(s-1),\alpha}$ es el cuantil de la distribución $\chi^2_{(r-1)(s-1)}$ que deja a su derecha una probabilidad igual a α.

$$\begin{aligned} q &= \frac{\left(20-50\,\frac{55}{110}\right)^2}{50\,\frac{55}{110}} + \frac{\left(20-50\,\frac{35}{110}\right)^2}{50\,\frac{35}{110}} + \frac{\left(10-50\,\frac{20}{110}\right)^2}{50\,\frac{20}{110}} + \\ &\quad + \frac{\left(35-60\,\frac{55}{110}\right)^2}{60\,\frac{55}{110}} + \frac{\left(15-60\,\frac{35}{110}\right)^2}{60\,\frac{35}{110}} + \frac{\left(10-60\,\frac{20}{110}\right)^2}{60\,\frac{20}{110}} = 3.9336, \end{aligned}$$

y como se tiene que el cuantil, $\chi^2_{(2-1)(3-1),0.01} = \chi^2_{2,0.01} = 9.21$, la decisión es ACEPTAR H_0 para un nivel de significación $\alpha = 0.01$ y, por consiguiente, se concluye que la distribución de los efectos producidos en los naranjos es la misma en ambos fertilizantes para dicho nivel de significación.

b) $H_0:\ \pi_{21} = \pi^*_{21} = 2\theta,\ \pi_{22} = \pi^*_{22} = \theta,\ \pi_{23} = \pi^*_{23} = 0.16$
$H_1:$ no H_0.
Como $2\theta + \theta + 0.16 = 1$, se tiene que $\theta = 0.28$ y no tenemos ningún parámetro desconocido en la hipótesis nula.
El estadístico χ^2 de Pearson cumple que

$$Q = \sum_{j=1}^{s} \frac{\left(n_{2j} - n\,\pi^*_{2j}\right)^2}{n\,\pi^*_{2j}} \mathrel{\dot\sim} \chi^2_{s-1}$$

y el test que resuelve el contraste de bondad de ajuste es rechazar $H_0 \Leftrightarrow q \geq \chi^2_{s-1,\alpha}$, con $\chi^2_{s-1,\alpha}$ el cuantil de la distribución χ^2_{s-1} que deja a su derecha una probabilidad igual a α.

$$q = \frac{(35 - 60 \times 0.56)^2}{60 \times 0.56} + \frac{(15 - 60 \times 0.28)^2}{60 \times 0.28} + \frac{(10 - 60 \times 0.16)^2}{60 \times 0.16} = 0.2679,$$

el cuantil $\chi^2_{2,0.8757} \simeq 0.2679$, luego la decisión es ACEPTAR H_0 para todo nivel de significación $\alpha \leq 0.8757$. Por consiguiente, sí que pueden aceptarse como válidos los porcentajes propuestos para los efectos del fertilizante F_2 para cualquier nivel de significación habitual.

5. a) Comparar los errores cuadráticos medios de los estimadores que las generan y quedarnos con la estimación proveniente del estimador con menor error cuadrático medio.

b) Sí, cabe la posibilidad de que no pueda establecerse una preferencia toda vez que el correspondiente error cuadrático medio dependa de algún parámetro desconocido.