# Examen de Estadística II. 28-01-2008.

**1.** $X$, producción de un avellano en kilogramos.
$\mu = 20$, $\sigma = 5$.

Como disponemos de una m.a.s. $(x_1, \ldots, x_{200})$ y el tamaño muestral $n = 200$ es suficientemente grande, estamos en condiciones de aplicar el teorema central del límite para obtener que $\sum_{i=1}^{n} X_i \dot\sim N(n\mu, n\sigma^2) = N(4000, 5000)$.

Podemos considerar el beneficio obtenido por la parcela en euros como una variable aleatoria $B = (k - 0.6) \sum_{i=1}^{200} X_i - 550$, donde $k$ es la constante requerida, el precio por kilogramo en euros. Al ser combinación lineal de una variable aproximadamente Normal se tiene que $B \dot\sim N(\mu_b = (k-0.6)4000 - 550, \sigma_b^2 = (k-0.6)^2(5000))$.

$$p\,(B > 1200) = 0.95 \quad \Leftrightarrow \quad p\left(Z > \frac{1200 - (k-0.6)4000 + 550}{(k-0.6)\sqrt{5000}}\right) = 0.95 \Leftrightarrow$$
$$\Leftrightarrow \quad \frac{4150 - 4000k}{(k-0.6)\sqrt{5000}} = -1.645 \Leftrightarrow \underline{k = 1.0506},$$

donde $Z = \frac{B - \mu_b}{\sigma_b} \dot\sim N(0,1)$.

**2.** $X$, igual a 1 si el niño recibe como máximo tres regalos y 0 en caso contrario.
$X \sim Br(\theta)$.
$(x_1, \ldots, x_{200})$ m.a.s. tal que $\sum_{i=1}^{200} x_i = 40$.

a) Como el tamaño muestral $n = 200$ es grande, el error aproximado $\epsilon$ asociado con la estimación de la proporción $\theta$ es $\epsilon = z_{\frac{\alpha}{2}} \sqrt{\frac{\overline{x}(1-\overline{x})}{n}}$, donde $1 - \alpha$ es la confianza requerida y $z_{\frac{\alpha}{2}}$ es el cuantil de la distribución $N(0,1)$ que deja a su derecha una probabilidad igual a $\frac{\alpha}{2}$. Por lo tanto, $\epsilon = z_{0.05} \sqrt{\frac{(0.2)(0.8)}{200}} = \underline{0.0465}$.

b) Como disponemos de una estimación previa de la varianza poblacional, $\hat\sigma^2 = s^2 = 0.16 = (0.2)(1-0.2)$, el tamaño muestral $n$ requerido para la estimación de la proporción $\theta$ con una determinada confianza $1 - \alpha$ y un error máximo $\epsilon$ puede ser obtenido mediante $n = z_{\frac{\alpha}{2}}^2 \frac{\hat\sigma^2}{\epsilon^2} = z_{0.025}^2 \frac{0.16}{(0.0465)^2} = 284.2669 \simeq \underline{285}$.

También puede ser propuesta la solución $n = z_{0.025}^2 \frac{0.25}{(0.0465)^2} = 443.6397 \simeq \underline{444}$, habiendo aproximado la varianza poblacional por su valor máximo.

**3.** a) Los dos investigadores tendrán la misma fiabilidad para con sus conclusiones dada la suficiencia del estadístico total para la inferencia sobre $\theta$ bajo el modelo $F(x, \theta)$.

b) Si el problema se plantea como un contraste de independencia entre las variables aleatorias medio de transporte utilizado y titulación cursada deberíamos seleccionar una muestra aleatoria simple de tamaño elevado del colectivo de estudiantes. Previa pregunta acerca del transporte utilizado y la titulación cursada quedarían clasificados según ese doble criterio.

Si se plantea como un contraste de homogeneidad entre las variables aleatorias obtenidas para las diferentes titulaciones al clasificar a los estudiantes según el medio de transporte utilizado deberíamos seleccionar una muestra aleatoria simple en cada una de las titulaciones bajo estudio de modo que la suma de todos los tamaños muestrales fuese elevada. La misma pregunta a realizar a cada colectivo sería obviamente el tipo de transporte utilizado quedando clasificados entonces según un único criterio.

**4.** $X$, número de pruebas de alcoholemia realizadas en un punto de control.
$X \sim Po(\lambda)$.
$(x_1, \ldots, x_{20})$ m.a.s. tal que $\sum_{i=1}^{20} x_i = 300$.

a) La media de una distribución Poisson resulta igual al parámetro de la misma, de modo que el estimador máximo verosímil de la media se obtiene maximizando la función de verosimilitud del parámetro $\lambda$ de la distribución Poisson,

$$l(\lambda) = \prod_{i=1}^{20} f(x_i; \lambda) = \prod_{i=1}^{20} \exp(-\lambda)\frac{\lambda^{x_i}}{x_i!} = \exp(-20\lambda)\frac{\lambda^{\sum_{i=1}^{20} x_i}}{\prod_{i=1}^{20} x_i!}.$$

De modo que la log-verosimilitud es $\log l(\lambda) = -20\lambda + \left(\sum_{i=1}^{20} x_i\right)\log\lambda - \log\prod_{i=1}^{20} x_i!$ y su primera derivada resulta $\frac{d \log l(\lambda)}{d\lambda} = -20 + \frac{\sum_{i=1}^{20} x_i}{\lambda}$, que igualada a cero proporciona la solución $\hat{\lambda} = \frac{\sum_{i=1}^{20} x_i}{20} = \overline{x} = 15$. Esta solución es efectivamente el máximo de $\log l(\lambda)$ (y por la inyectividad de la función logarítmica, también de $l(\lambda)$), pues hace negativa su segunda derivada, $\frac{d^2 \log l(\lambda)}{d\lambda^2} = -\frac{\sum_{i=1}^{20} x_i}{\lambda^2} < 0$, $\forall \lambda > 0$, luego la media muestral $\overline{X}$ es el estimador máximo verosímil de la media $\lambda$ de una población Poisson y, en este caso, $\underline{\overline{x} = 15}$ es la estimación máximo verosímil.

b) Para comprobar la eficiencia de $\overline{X}$ hemos de verificar que es un estimador insesgado para $\mu = \lambda$ cuya varianza alcanza la cota de Kramer-Rao. La insesgadez es inmediata dado que la media muestral resulta siempre un estimador insesgado de la media poblacional, $E\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n} n\mu = \mu$.

La cota de Kramer-Rao es igual al inverso de la información de Fisher que podemos obtener como $I(\lambda) = -E\left[\frac{d^2 \log f(x_1, \ldots x_n; \lambda)}{d\lambda^2}\right]$. Para el caso Poisson que nos ocupa adquiriría la forma $I(\lambda) = -E\left[-\frac{\sum_{i=1}^{20} x_i}{\lambda^2}\right] = \frac{20\lambda}{\lambda^2} = \frac{20}{\lambda}$. Y efectivamente su inverso, $\frac{\lambda}{20}$, coincide con la varianza de la media muestral toda vez que ésta es igual a la varianza poblacional entre el tamaño muestral, $Var\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(x_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$.

**5.** $X$, destino del tranvía.
$Y$, parada del tranvía.

$H_0$ : $X$ e $Y$ independientes
$H_1$ : no $H_0$.

Bajo ciertas condiciones genéricas y para tamaños muestrales grandes se tiene que

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

El test que resuelve el contraste de independencia es rechazar $H_0 \Leftrightarrow q \geq \chi^2_{(r-1)(s-1),\alpha}$, 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 $\alpha$.

$$\begin{aligned} q &= \frac{\left(3 - \frac{(75)(51)}{200}\right)^2}{\frac{(75)(51)}{200}} + \frac{\left(60 - \frac{(75)(109)}{200}\right)^2}{\frac{(75)(109)}{200}} + \frac{\left(12 - \frac{(75)(40)}{200}\right)^2}{\frac{(75)(40)}{200}} + \\ &+ \frac{\left(33 - \frac{(82)(51)}{200}\right)^2}{\frac{(82)(51)}{200}} + \frac{\left(31 - \frac{(82)(109)}{200}\right)^2}{\frac{(82)(109)}{200}} + \frac{\left(18 - \frac{(82)(40)}{200}\right)^2}{\frac{(82)(40)}{200}} + \\ &+ \frac{\left(15 - \frac{(43)(51)}{200}\right)^2}{\frac{(43)(51)}{200}} + \frac{\left(18 - \frac{(43)(109)}{200}\right)^2}{\frac{(43)(109)}{200}} + \frac{\left(10 - \frac{(43)(40)}{200}\right)^2}{\frac{(43)(40)}{200}} = 37.4573, \end{aligned}$$

y como se tiene que el cuantil, $\chi^2_{(3-1)(3-1),0.005} = \chi^2_{4,0.005} = 14.86 < 37.4573 = q$, la decisión es RECHAZAR $H_0$ para un nivel de significación del 0.5% y evidentemente también para cualquiera de los habituales. Consecuentemente, se concluye que el destino del tranvía elegido por el usuario sí que está relacionado con la parada utilizada.

**6.** $X$, importe en euros de un viaje a Europa.
$(x_1, \ldots, x_{15})$ m.a.s. tal que $\overline{x} = 1250, s = 174$.

Como la muestra no puede ser considerada de tamaño elevado se tiene que suponer Normalidad para la población, esto es, $X \sim N(\mu, \sigma^2)$, de donde $\sqrt{n-1}\, \frac{\overline{X}-\mu}{S} \sim t_{n-1}$.

Hemos de hallar el nivel de significación crítico del contraste de hipótesis

$$\begin{aligned} H_0 &: \mu \leq \mu_0 = 1200 \\ H_1 &: \mu > \mu_0 = 1200. \end{aligned}$$

El test con un nivel de significación $\alpha$ para su resolución es

$$\text{rechazar } H_0 \Leftrightarrow \overline{x} \geq \mu_0 + t_{n-1,\alpha}\, \frac{s}{\sqrt{n-1}}.$$

Resolviendo la ecuación para $\alpha$, $\overline{x} = \mu_0 + t_{n-1,\alpha}\, \frac{s}{\sqrt{n-1}} \Leftrightarrow 1250 = 1200 + t_{14,\alpha}\, \frac{174}{\sqrt{14}}$, obtenemos $t_{14,\alpha} = 1.0752$ y el nivel de significación crítico $\alpha_0 = 0.15$.