Examen de Estadística II. 02-02-2010.

1. X, número de vehículos multados un sábado.
Y, número de vehículos multados un domingo.
$(x_1, \ldots, x_{52})$ m.a.s. tal que $\mu_x = 30$, $\sigma_x = 5$.
$(y_1, \ldots, y_{52})$ m.a.s. tal que $\mu_y = 45$, $\sigma_y = 10$.

Asumiendo que el número de vehículos multados los sábados son variables aleatorias independientes y equidistribuidas podemos considerar $(x_1, \ldots, x_{52})$ como una m.a.s. Además, como el tamaño muestral $n = 52$ es suficientemente grande estamos en condiciones de aplicar el teorema central del límite para obtener que $\sum_{i=1}^{n} X_i \dot{\sim} N(n\mu_x, n\sigma_x^2) = N(1560, 1300)$. De un modo totalmente análogo, $\sum_{i=1}^{n} Y_i \dot{\sim} N(n\mu_y, n\sigma_y^2) = N(2340, 5200)$.

Podemos considerar la recaudación anual en euros por multas en fines de semana como la variable aleatoria $W = 40 \sum_{i=1}^{52} (X_i + Y_i)$. Asumiendo independencia entre las variables X e Y se tiene que $W \dot{\sim} N(\mu_w = 40(1560 + 2340) = 156000, \sigma_w^2 = 40^2(1300 + 5200) = 10400000)$, al ser combinación lineal de variables aproximadamente Normales.

$$p\left(W > 150000\right) = p\left(Z > \frac{150000 - 156000}{\sqrt{10400000}}\right) = p(Z > -1.8605) = p(Z < 1.8605) = \underline{0.9686},$$

donde $Z = \frac{W - \mu_w}{\sigma_w} \dot{\sim} N(0, 1)$.

2. X, tiempo de servicio de un producto en días.
$X \sim Ex(\theta)$, $f(x) = \frac{1}{\theta} \exp\left(-\frac{1}{\theta} x\right)$, $x > 0$, $\theta > 0$.
$(x_1, \ldots, x_{50})$ m.a.s. tal que $\sum_{i=1}^{50} x_i = 225$.

El estimador máximo verosímil de θ se obtiene maximizando la función de verosimilitud Exponencial,

$$l(\theta) = \prod_{i=1}^{50} f(x_i; \theta) = \prod_{i=1}^{50} \frac{1}{\theta} \exp\left(-\frac{1}{\theta} x_i\right) = \frac{1}{\theta^{50}} \exp\left(-\frac{1}{\theta} \sum_{i=1}^{50} x_i\right).$$

De modo que la log-verosimilitud es $\log l(\theta) = -50 \log \theta - \frac{1}{\theta}\left(\sum_{i=1}^{50} x_i\right)$ y su primera derivada resulta $\frac{d \log l(\theta)}{d\theta} = -\frac{50}{\theta} + \frac{\sum_{i=1}^{25} x_i}{\theta^2}$, que igualada a cero proporciona la solución $\hat{\theta} = \frac{\sum_{i=1}^{50} x_i}{50} = \overline{x}$. Efectivamente, 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, $\frac{d^2 \log l(\theta)}{d\theta^2}\Big|_{\theta = \overline{x}} = \frac{50}{\overline{x}^2} - \frac{2\sum_{i=1}^{50} x_i}{\overline{x}^3} = \frac{50}{\overline{x}^2} - \frac{100}{\overline{x}^2} = -\frac{50}{\overline{x}^2} < 0$, luego la media muestral $\overline{X}$ es el estimador máximo verosímil de la media $\mu = \theta$ de una población Exponencial y, en este caso, $\overline{x} = \frac{\sum_{i=1}^{50} x_i}{50} = \underline{4.5}$ es la estimación máximo verosímil.

3. X_1, igual a 1 si un votante andaluz tiene intención de votar al partido, 0 en caso contrario.
X_2, igual a 1 si un votante valenciano tiene intención de votar al partido, 0 en caso contrario.

a) Un posible plan de muestreo consistiría en seleccionar sendas muestras aleatorias simples de ambas poblaciones Bernouilli. Utilizando, por ejemplo, el censo electoral andaluz podríamos seleccionar uno de cada 50 electores y preguntarle su intención de voto al partido hasta completar una m.a.s., $(x_{11},\ldots,x_{n_1 1})$, de tamaño elevado (mayor que 30, por ejemplo). Procederíamos de un modo completamente análogo con el censo valenciano para la obtención de una m.a.s. $(x_{12},\ldots,x_{n_2 2})$.

b) Los estadísticos muestrales a utilizar serían las respectivas medias o proporciones muestrales, $\overline{x}_1$ y $\overline{x}_2$.

c) Asumiendo la obvia independencia entre X_1 y X_2 y aplicando el teorema central del límite, el test con un nivel de significación $\alpha = 0.05$ que se propone es rechazar $H_0 \Leftrightarrow |\overline{x}_1 - \overline{x}_2| > (0.5)\, z_{0.025}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = (0.98)\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$, siendo H_0 la hipótesis nula de igualdad de proporciones poblacionales y $z_{0.025} = 1.96$ el cuantil de la distribución $N(0,1)$ que deja a su derecha una probabilidad igual a 0.025. La correspondiente región crítica o región de rechazo es $C_1 = \left\{(x_{11},\ldots,x_{n_1 1}),\ (x_{12},\ldots,x_{n_2 2}),\ \text{t.q. } |\overline{x}_1 - \overline{x}_2| > (0.98)\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right\}$.

O bien el test alternativo con un riesgo de tipo I más ajustado a $\alpha = 0.05$, rechazar $H_0 \Leftrightarrow |\overline{x}_1 - \overline{x}_2| > (1.96)\sqrt{\hat{\theta}(1-\hat{\theta})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ donde $\widehat{\theta} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2}{n_1+n_2}$, combinación lineal convexa de las respectivas medias muestrales, es el estimador máximo verosímil de la proporción poblacional desconocida obtenido con el conjunto de datos provenientes de las dos muestras y asumiendo la igualdad de proporciones poblacionales que especifica la hipótesis nula. En este caso la región crítica es $C_2 = \left\{(x_{11},\ldots,x_{n_1 1}),\ (x_{12},\ldots,x_{n_2 2}),\ \text{t.q. } |\overline{x}_1 - \overline{x}_2| > (1.96)\sqrt{\hat{\theta}(1-\hat{\theta})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right\}$.

4. X, igual a 1 si la pieza es defectuosa y 0 en caso contrario.
$X \sim Br(\theta)$.

a) $1-\alpha = 0.95,\ \epsilon = 0.05$.

El tamaño muestral n requerido para la estimación de una proporción con una determinada confianza $1-\alpha$ y un error máximo ϵ es $n = z^2_{\frac{\alpha}{2}}\frac{0.25}{\epsilon^2}$, donde $z_{\frac{\alpha}{2}}$ es el cuantil de la distribución $N(0,1)$ que deja a su derecha una probabilidad igual a $\frac{\alpha}{2}$. De este modo, el tamaño necesario para la estimación de θ con la precisión requerida es $n = z^2_{0.025}\frac{0.25}{(0.05)^2} = (1.96)^2\frac{0.25}{(0.05)^2} = 384.16$. Consecuentemente, deberán ser examinadas $n = 385$ piezas para asegurar la precisión de la estimación.

b) $(x_1,\ldots,x_{100})$ m.a.s. tal que $\sum_{i=1}^{100} x_i = 4$.

Se ha de resolver el siguiente contraste de hipótesis

$$\begin{aligned} H_0 &: \theta \le \theta_0 = \frac{100}{3000} = 0.0\widehat{3} \\ H_1 &: \theta > \theta_0 = 0.0\widehat{3}. \end{aligned}$$

Como el tamaño muestral $n = 100$ es suficientemente grande, podemos aplicar el teorema central del límite para obtener que $\overline{X} \sim N\left(\theta, \frac{\theta(1-\theta)}{n}\right)$, de modo que el test con un nivel de significación α que se propone para resolver el contraste planteado es

$$\text{rechazar } H_0 \Leftrightarrow \overline{x} \geq \theta_0 + z_\alpha \sqrt{\frac{\theta_0(1-\theta_0)}{n}}.$$

Hallamos el nivel de significación crítico resolviendo la ecuación asociada a la inecuación anterior, $\frac{4}{100} = 0.04 = 0.0\widehat{3} + z_\alpha \sqrt{\frac{(0.0\widehat{3})(1-0.0\widehat{3})}{100}}$, de donde $z_\alpha = 0.3714$ y la decisión es ACEPTAR $H_0 \; \forall \, \alpha \leq 0.3552$, es decir, en base a los resultados muestrales debe aceptarse el lote para todo nivel de significación habitual.

5. X, destino vacacional preferido por un extranjero.

$H_0: \; \pi_1 = \pi_1^* = 0.25, \; \pi_2 = \pi_2^* = 0.25, \; \pi_3 = \pi_3^* = 0.15, \; \pi_4 = \pi_4^* = 0.15, \; \pi_5 = \pi_5^* = \theta$
$H_1: \;$ no H_0.

Como $0.25+0.25+0.15+0.15+\theta = 1$, se tiene que $\theta = 0.2$ y no tenemos ningún parámetro desconocido en la hipótesis nula.

El estadístico χ^2 de Pearson cumple que

$$Q = \sum_{i=1}^{r} \frac{(n_i - n\,\pi_i^*)^2}{n\,\pi_i^*} \dot{\sim} \chi^2_{r-1}$$

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

$$q = \frac{(120 - 400 \times 0.25)^2}{400 \times 0.25} + \frac{(112 - 400 \times 0.25)^2}{400 \times 0.25} + \frac{(60 - 400 \times 0.15)^2}{400 \times 0.15} + \frac{(48 - 400 \times 0.15)^2}{400 \times 0.15} +$$

$$+ \frac{(60 - 400 \times 0.2)^2}{400 \times 0.2} = 12.84,$$

el cuantil $\chi^2_{4,0.0131} \simeq 12.84$, luego la decisión es ACEPTAR H_0 para todo nivel de significación $\alpha \leq 0.0131$. Por consiguiente, sólo puede aceptarse la compatibilidad entre los resultados de la encuesta y las previsiones publicadas para todo el año para el nivel de significación habitual $\alpha = 0.01$.

6. X_1, igual a 1 si la familia de la ciudad A tiene dificultades para llegar a fin de mes, 0 en caso contrario.

X_2, igual a 1 si la familia de la ciudad B tiene dificultades para llegar a fin de mes, 0 en caso contrario.

$X_1 \sim Br(\theta_1)$, $(x_{11}, \ldots, x_{200\,1})\, m.a.s.$, $\epsilon_1 = 0.03$, $1 - \alpha_1 = 0.9$.
$X_2 \sim Br(\theta_2)$, $(x_{12}, \ldots, x_{150\,2})\, m.a.s.$, $\epsilon_2 = 0.04$, $1 - \alpha_2 = 0.95$.

Para poder comparar las precisiones es necesario igualar los niveles de confianza. Por ejemplo, podemos estimar la varianza de X_1, $\hat{\sigma}_1^2 = \frac{n_1\,\epsilon_1^2}{z^2_{\frac{\alpha_1}{2}}} = 200\frac{0.03^2}{1.645^2} = 0.0665$ y calcular el error asociado a la estimación de θ_1 con dicha aproximación de la varianza y con el mismo nivel de confianza $1 - \alpha_2 = 0.95$, $\epsilon = z_{\frac{\alpha_2}{2}}\sqrt{\frac{\hat{\sigma_1^2}}{n_1}} = (1.96)\sqrt{\frac{0.0665}{200}} = 0.0357 < 0.04 = \epsilon_2$ y, por consiguiente, el analista A ofrece una mayor precisión para su resultado.