Examen de Estadística II. 29-01-2002.

1. X, importe I.P.F. en miles de euros constituido en Patagon.
Y, importe I.P.F. en miles de euros constituido en ING Direct.
$\mu_x = 2,\ \sigma_x = 1,\ \mu_y = 3,\ \sigma_y = 1.5.$

$B = 0.01\left(\sum_{i=1}^{50} X_i + \sum_{i=1}^{80} Y_i\right)$, beneficio en miles de euros obtenido por el gestor.

Asumiendo que no existe relación entre los diferentes depósitos constituidos por el agente en Patagon y que todos ellos tienen la misma media y la misma varianza, como el tamaño muestral $n_x = 50$ es suficientemente grande, estamos en condiciones de aplicar el teorema central del límite para obtener que $\sum_{i=1}^{50} X_i \dot{\sim} N(50 \cdot 2, 50 \cdot 1^2) = N(100, 50)$.

Razonando de un modo totalmente análogo, $\sum_{i=1}^{80} Y_i \dot{\sim} N(80 \cdot 3, 80\,(1.5)^2) = N(240, 180)$. Así, como el beneficio es una combinación lineal de variables aleatorias Normales, su distribución de probabilidad también es asintóticamente Normal, con media $\mu_B = 0.01(100 + 240) = 3.4$ y varianza $\sigma_B^2 = (0.01)^2(50 + 180) = 0.023$, pues existe independencia entre X e Y.

De este modo, calculamos de forma aproximada la probabilidad requerida como:

$$p(B > 3) = p\left(Z > \frac{3 - 3.4}{\sqrt{0.023}}\right) = p(Z > 2.6375) = \underline{0.9959},$$

con $Z = \frac{B - \mu_B}{\sigma_B} \dot{\sim} N(0, 1)$.

2. X, número de llamadas diarias recibidas en el parque de bomberos.
$X \sim Po(\lambda),\ f(x; \lambda) = \exp(-\lambda)\frac{\lambda^x}{x!},\ x > 0,\ \lambda > 0.$
$(x_1, \ldots, x_{20})$ m.a.s.

a) Como las variables aleatorias asociadas con la muestra son Poisson independientes, la función de verosimilitud resulta:

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

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 $\hat{\lambda} = \frac{\sum_{i=1}^{20} x_i}{20} = \overline{x}$. 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 λ de una población Poisson, c.q.d.

b) Un estimador es insesgado cuando su sesgo es cero, es decir, cuando su media coincide con el parámetro que pretende estimar, cualquiera que sea el valor de éste. Por lo tanto, hemos de demostrar que $E\left(\overline{X}\right)=\lambda,\ \forall\lambda>0$.

$E\left(\overline{X}\right)=E\left(\frac{\sum_{i=1}^{20}X_i}{20}\right)=\frac{1}{20}\sum_{i=1}^{20}E(X_i)=\frac{1}{20}\sum_{i=1}^{20}\lambda=\frac{20\lambda}{20}=\lambda,\ \forall\lambda>0$, c.q.d.

c) El error cuadrático medio de un estimador es igual a su varianza más el cuadrado del sesgo, por lo que para estimadores insesgados resulta igual a la varianza del estimador.

$Var\left(\overline{X}\right)=Var\left(\frac{\sum_{i=1}^{20}X_i}{20}\right)=\frac{1}{20^2}\sum_{i=1}^{20}Var(X_i)=\frac{1}{20^2}\sum_{i=1}^{20}\lambda=\frac{20\lambda}{20^2}=\frac{\lambda}{20}$.

d) Sí, comprobando que el estimador es eficiente, esto es, que además de insesgado su varianza alcanza la cota de Cramer-Rao. De este modo, su varianza sería la mínima entre la de todos los estimadores insesgados y consecuentemente el estimador sería insesgado óptimo.

Como ya se ha demostrado la insesgadez, sólo habría que comprobar la condición acerca de la varianza, es decir, que $Var\left(\overline{X}\right)=\frac{1}{I(\lambda)},\ \forall\lambda>0$, donde $I(\lambda)=E\left[\left(\frac{d\log f(X_1,\ldots,X_{20};\lambda)}{d\lambda}\right)^2\right]=-E\left[\frac{d^2\log f(X_1,\ldots,X_{20};\lambda)}{d\lambda^2}\right]$.

3. X_1, igual a 1 si el elector andaluz vota al partido y 0 en caso contrario.
X_2, igual a 1 si el elector valenciano vota al partido y 0 en caso contrario.
$X_1\sim Br(\theta_1),\ X_2\sim Br(\theta_2)$.

a) Debería ser seleccionada una muestra aleatoria simple de X_1, de tamaño n_1 elevado, así como una muestra aleatoria simple de X_2, también de tamaño n_2 suficientemente grande.

Para contrastar la opinión del líder político, se ha de resolver el contraste de hipótesis:

$$\begin{aligned} H_0 &: \theta_1\geq\theta_2\\ H_1 &: \theta_1<\theta_2\,. \end{aligned}$$

Asumiendo independencia entre las dos poblaciones X_1 y X_2 y aplicando el teorema central del límite, se tiene que $\overline{X}_1-\overline{X}_2\dot{\sim}N\left(\theta_1-\theta_2,\frac{\theta_1(1-\theta_1)}{n_1}+\frac{\theta_2(1-\theta_2)}{n_2}\right)$, de modo que el test con un nivel de significación $\alpha=0.05$ que se propone para resolver el contraste planteado es:

$$\text{rechazar } H_0\Leftrightarrow\overline{x}_1-\overline{x}_2\leq-(0.5)\ z_{0.05}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}=-(0.8225)\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},$$

o bien el test alternativo con un riesgo de tipo I más ajustado a $\alpha=0.05$:

$$\text{rechazar } H_0\Leftrightarrow\overline{x}_1-\overline{x}_2\leq-(1.645)\sqrt{\hat{\theta}(1-\hat{\theta})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)},$$

donde $z_{0.05}=1.645$ es el cuantil de la distribución $N(0,1)$ que deja a su derecha una probabilidad igual a 0.05 y $\widehat{\theta}=\frac{n_1\overline{x}_1+n_2\overline{x}_2}{n_1+n_2}$, 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.

b) El nivel de significación crítico es el nivel hasta el cual se acepta la hipótesis nula y a partir del cual se rechaza. Al ser en este caso del 1%, lo razonable sería rechazar H_0. De hecho, para cualquiera de los niveles habituales la decisión correcta pasa por rechazar la hipótesis nula, sólo para niveles de significación estrictamente inferiores a 0.01 podríamos aceptar H_0.

4. X, igual a 1 si el agricultor asegura la cosecha y 0 en caso contrario.
$X \sim Br(\theta)$.

a) $\epsilon = 0.04,\ 1-\alpha = 0.95$.
Para determinar el tamaño muestral necesario para alcanzar unos determinados requerimientos en cuanto a la precisión de una estimación de la proporción θ, utilizamos el hecho de que la media de una m.a.s. de tamaño grande sigue aproximadamente una distribución Normal de media θ y varianza $\frac{\theta(1-\theta)}{n}$, en virtud del teorema central del límite.
Así, $n = (z_{\frac{\alpha}{2}})^2\ \frac{\theta(1-\theta)}{\epsilon^2} = (1.96)^2\frac{0.25}{(0.04)^2} = 600.25$, donde se ha sustituido la varianza poblacional desconocida $\theta(1-\theta)$ por su valor máximo, alcanzado en $\theta = 0.5$.
De este modo, es necesario entrevistar a 601 agricultores.

b) Y, prima anual por agricultor en euros.
$Y \sim N(\mu, \sigma^2)$.
$(y_1, \dots, y_{101})$ m.a.s. tal que $\overline{x} = 120,\ s = 40$.
Una buena estimación para μ, la prima media poblacional desconocida, es $\overline{x} = 120$, pues bajo un modelo Normal la media muestral es el e.m.v. de la media poblacional, es insesgado, suficiente, consistente y eficiente.

c) $1-\alpha = 0.9$.
Como se tiene que la población es Normal, es de aplicación el teorema de Cochran que establece que $\frac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$. A partir de este resultado, se obtiene el intervalo de confianza $1-\alpha$ para σ^2 como $\left[\frac{ns^2}{\chi^2_{n-1,\frac{\alpha}{2}}}, \frac{ns^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right]$, donde $\chi^2_{n-1,\frac{\alpha}{2}}$ y $\chi^2_{n-1,1-\frac{\alpha}{2}}$ son los cuantiles de la distribución χ^2_{n-1} que dejan a su derecha probabilidades iguales a $\frac{\alpha}{2}$ y $\left(1-\frac{\alpha}{2}\right)$, respectivamente.
Utilizando la información muestral proporcionada en el apartado anterior y consultando en tablas los cuantiles $\chi^2_{100,0.05} = 124.3$ y $\chi^2_{100,0.95} = 77.93$, el intervalo numérico de confianza del 90% para σ^2 resulta $\left[\frac{101(40)^2}{\chi^2_{100,0.05}}, \frac{101(40)^2}{\chi^2_{100,0.95}}\right] = \underline{[1300.0805, 2073.6558]}$.

5. X, marca elegida por el cliente que compra el producto.
Y, grado de satisfacción del cliente con el producto.

a) Hemos de resolver el contraste de independencia:

$$\begin{aligned} H_0 &: X \text{ e } Y \text{ independientes} \\ H_1 &: \text{no } H_0\,. \end{aligned}$$

Bajo ciertas condiciones genéricas y para un tamaño muestral grande 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 planteado es:

$$\text{rechazar } H_0 \Leftrightarrow q \geq \chi^2_{(r-1)(s-1),\alpha},$$

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 α.

En nuestro caso, con $r=2$ categorías para X y $s=3$ categorías para Y,

$$\begin{aligned} q \;=\; & \frac{\left(35-\frac{(120)(47)}{200}\right)^2}{\frac{(120)(47)}{200}}+\frac{\left(70-\frac{(120)(120)}{200}\right)^2}{\frac{(120)(120)}{200}}+\frac{\left(15-\frac{(120)(33)}{200}\right)^2}{\frac{(120)(33)}{200}}+ \\ & +\frac{\left(12-\frac{(80)(47)}{200}\right)^2}{\frac{(80)(47)}{200}}+\frac{\left(50-\frac{(80)(120)}{200}\right)^2}{\frac{(80)(120)}{200}}+\frac{\left(18-\frac{(80)(33)}{200}\right)^2}{\frac{(80)(33)}{200}}=7.1473. \end{aligned}$$

Y como se tiene que el estadístico es mayor que el cuantil, $q=7.1473>5.991=\chi^2_{2,0.05}$, la decisión es <u>RECHAZAR H_0</u> para un nivel de significación $\alpha=0.05$. Se ha de ser prudente con la conclusión pues si se hubiera considerado un nivel de significación algo más conservador, por ejemplo $\alpha=0.01$, entonces $\chi^2_{2,0.01}=9.21$ y la decisión correcta habría sido que la marca con la que se comercializa el producto no influye en el grado de satisfacción del cliente, para dicho nivel de significación.

b) En este caso se trata de un contraste de bondad de ajuste.

$$\begin{aligned} H_0 &: \;\pi_{\cdot 1}=\pi^*_{\cdot 1}=\theta,\; \pi_{\cdot 2}=\pi^*_{\cdot 2}=2\theta,\; \pi_{\cdot 3}=\pi^*_{\cdot 3}=0.15 \\ H_1 &: \;\text{no } H_0\,, \end{aligned}$$

donde $\theta+2\theta+0.15=1$, es decir, $\theta=0.28\hat{3}$ y no tenemos ningún parámetro desconocido en la hipótesis nula.

El estadístico χ^2 de Pearson cumple que:

$$Q=\sum_{j=1}^{3}\frac{\left(n_{\cdot j}-n\,\pi^*_{\cdot j}\right)^2}{n\,\pi^*_{\cdot j}}\,\dot{\sim}\,\chi^2_{s-1}\,,$$

y el test que resuelve el contraste de bondad de ajuste es:

$$\text{rechazar } H_0 \Leftrightarrow q \geq \chi^2_{s-1,\,\alpha},$$

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.

$$q=\frac{\left(47-200\,(0.28\hat{3})\right)^2}{200\,(0.28\hat{3})}+\frac{\left(120-200\,(0.5\hat{6})\right)^2}{200\,(0.5\hat{6})}+\frac{\left(200\,(0.15)\right)^2}{200\,(0.15)}=2.3412,$$

que es menor que el cuantil $\chi^2_{2,0.1}=4.605$, luego la decisión es <u>ACEPTAR H_0</u> para un nivel de significación $\alpha=0.1$.