Examen de Estadística II. 24-01-2003.

1. X_1, igual a 1 si el elector andaluz tiene intención de votar al PSOE y 0 en caso contrario. $X_1 \sim Br(\theta_1)$.
X_2, igual a 1 si el elector valenciano tiene intención de votar al PSOE y 0 en caso contrario. $X_2 \sim Br(\theta_2)$.

a) $\epsilon = 0.02,\ 1-\alpha = 0.95$.

Para determinar el tamaño muestral necesario para alcanzar unos determinados requerimientos en cuanto a la precisión de una estimación de la proporción θ_1, utilizamos el hecho de que la media de una m.a.s. de tamaño n_1 grande sigue aproximadamente una distribución Normal de media θ_1 y varianza $\frac{\theta_1(1-\theta_1)}{n_1}$, en virtud del teorema central del límite.

Así, $n_1 = (z_{\frac{\alpha}{2}})^2\ \frac{\theta_1(1-\theta_1)}{\epsilon^2} = (1.96)^2 \frac{0.25}{(0.02)^2} = 2401$, donde se ha sustituido la varianza poblacional desconocida $\theta_1(1-\theta_1)$ por su valor máximo, alcanzado en $\theta_1 = 0.5$.

De este modo, como este razonamiento también es perfectamente válido para n_2, es necesaria una muestra de tamaño <u>2401</u> en cada comunidad para alcanzar los objetivos de precisión requeridos.

b) $\epsilon = 0.02,\ 1-\alpha = 0.95,\ n_1 = n_2 = n$.

Combinando el resultado comentado en el apartado anterior para la distribución de la media muestral, con la propiedad reproductiva de la distribución Normal para transformaciones lineales, se tiene que la diferencia de las respectivas medias muestrales sigue asintóticamente una distribución Normal de media $\theta_1 - \theta_2$ y varianza $\frac{\theta_1(1-\theta_1)}{n_1} + \frac{\theta_2(1-\theta_2)}{n_2}$. Como los tamaños muestrales son iguales, el error de estimación cometido al aproximar la diferencia de proporciones poblacionales mediante la correspondiente diferencia de proporciones muestrales es $\epsilon = \sqrt{\frac{\theta_1(1-\theta_1)+\theta_2(1-\theta_2)}{n}}$.

Así, despejando el tamaño muestral requerido, se tiene que $n = (z_{\frac{\alpha}{2}})^2\ \frac{\theta_1(1-\theta_1)+\theta_2(1-\theta_2)}{\epsilon^2} = (1.96)^2 \frac{(2)(0.25)}{(0.02)^2} = 4802$, donde se han sustituido las varianzas poblacionales desconocidas $\theta_1(1-\theta_1)$ y $\theta_2(1-\theta_2)$, por sus valores máximos que se alcanzan en $\theta_1 = \theta_2 = 0.5$.

Son necesarias muestras de tamaño <u>4802</u> para alcanzar los objetivos de precisión requeridos.

2. X_1, gasto de una familia madrileña en euros en las Navidades de 2002.
X_2, gasto de una familia valenciana en euros en las Navidades de 2002.
$(x_1^1, \ldots, x_{300}^1)$ m.a.s. tal que $\overline{x}_1 = 390,\ s_1 = 120$.
$(x_1^2, \ldots, x_{250}^2)$ m.a.s. tal que $\overline{x}_2 = 295,\ s_2 = 100$.

a) Denotando por μ_1 y μ_2 a las respectivas medias poblacionales y mediante σ_1^2 y σ_2^2 a las varianzas, se ha de resolver el siguiente contraste de hipótesis

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 \leq \mu_0 = 80 \\ H_1 &: \mu_1 - \mu_2 > \mu_0 = 80. \end{aligned}$$

Aplicando el teorema central del límite, se tiene que $\overline{X}_1 - \overline{X}_2 \dot{\sim} N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2}\right)$, de modo que el test con un nivel de significación aproximado α que se propone para resolver el contraste planteado es

$$\text{rechazar } H_0 \Leftrightarrow \overline{x}_1 - \overline{x}_2 \geq \mu_0 + z_\alpha \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

donde, como ambos tamaños muestrales son grandes, se han sustituido las varianzas poblacionales desconocidas por sus correspondientes estimaciones, s_1^2 y s_2^2, las varianzas muestrales.

Hallamos el nivel de significación crítico α_0, nivel hasta el cual se acepta la hipótesis nula y a partir del cual se rechaza, resolviendo la ecuación $\overline{x}_1 - \overline{x}_2 = 390 - 295 = 80 + z_\alpha \sqrt{\frac{120^2}{300} + \frac{100^2}{250}} = \mu_0 + z_\alpha \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, de donde, $z_\alpha = 1.599$ y $\alpha_0 = 0.0548$.

Por lo tanto, al estar comprendido α_0 entre niveles habituales, la decisión ha de ser tomada con cierta cautela. En este caso, por ejemplo, para un nivel de significación $\alpha = 0.05$ la decisión es ACEPTAR H_0, mientras que para un nivel de significación $\alpha = 0.1$ la decisión es RECHAZAR H_0 y para dicho nivel no puede afirmarse con rigor científico que la diferencia entre los gastos medios no ha aumentado en relación a las Navidades anteriores.

b) El test con un nivel de significación α que resuelve el contraste de significación

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = \mu_0 = 0 \\ H_1 &: \mu_1 - \mu_2 \neq \mu_0 = 0 \end{aligned}$$

es rechazar $H_0 \Leftrightarrow \mu_0 \notin \left[(\overline{x}_1 - \overline{x}_2) - z_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, (\overline{x}_1 - \overline{x}_2) + z_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right]$.

Si las medias muestrales son iguales, $\overline{x}_1 - \overline{x}_2 = 0$, el nivel de significación crítico se halla resolviendo la ecuación $0 = -z_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = z_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, de donde $z_{\frac{\alpha}{2}} = 0$ y $\alpha_0 = 1$.

3. Asumiendo que la población sigue una distribución $N(\mu, \sigma^2)$ y dada una muestra aleatoria simple de ella de tamaño n, entonces el teorema de Cochran establece que $\frac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$. A partir de este resultado y suponiendo media poblacional μ desconocida, 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.

Como el intervalo de confianza de una transformación biyectiva de un parámetro θ es igual a la biyección del intervalo de confianza (de sus extremos) de θ y $g(\sigma^2) = \frac{1}{\sigma^2}$ es una biyección, entones el intervalo de confianza $1-\alpha$ para η resulta $\left[\frac{\chi^2_{n-1,1-\frac{\alpha}{2}}}{ns^2}, \frac{\chi^2_{n-1,\frac{\alpha}{2}}}{ns^2}\right]$, donde la única hipótesis que es necesario considerar es la Normalidad para la población.

En el caso de que la media poblacional μ fuera conocida entonces el intervalo de confianza $1-\alpha$ para η es $\left[\frac{\chi^2_{n,1-\frac{\alpha}{2}}}{\sum_{i=1}^n (x_i-\mu)^2}, \frac{\chi^2_{n,\frac{\alpha}{2}}}{\sum_{i=1}^n (x_i-\mu)^2}\right]$.

4. X, número mensual de personas que emigran de la provincia A a la provincia B.
$X \sim Po(\lambda)$.
$(x_1, \ldots, x_{36})$ m.a.s. tal que $\sum_{i=1}^{36} x_i = 396$.

a) La mejor estimación posible de un parámetro poblacional es la proporcionada por el estimador máximo verosímil que, como las variables aleatorias asociadas con la muestra se suponen Poisson independientes, se obtiene maximizando la función de verosimilitud Poisson.

$$l(\lambda) = \prod_{i=1}^{36} f(x_i; \lambda) = \prod_{i=1}^{36} \exp(-\lambda) \frac{\lambda^{x_i}}{x_i!} = \exp(-36\lambda) \frac{\lambda^{\sum_{i=1}^{36} x_i}}{\prod_{i=1}^{36} x_i!}.$$

De modo que la log-verosimilitud es $\log l(\lambda) = -36\lambda + \left(\sum_{i=1}^{36} x_i\right) \log \lambda - \log \prod_{i=1}^{36} x_i!$ y su primera derivada resulta $\frac{d \log l(\lambda)}{d\lambda} = -36 + \frac{\sum_{i=1}^{36} x_i}{\lambda}$, que igualada a cero proporciona la solución $\hat{\lambda} = \frac{\sum_{i=1}^{36} x_i}{36} = \underline{\overline{x} = 11}$. 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, $\frac{d^2 \log l(\lambda)}{d\lambda^2} = -\frac{\sum_{i=1}^{36} x_i}{\lambda^2} < 0$, $\forall \lambda > 0$, luego la media muestral $\overline{X}$ es el estimador máximo verosímil de la media λ de una población Poisson, c.q.d.

b) Asumiendo que no existe relación entre los diferentes movimientos migratorios mensuales y que todos ellos tienen la misma distribución que X, como el tamaño muestral $n = 120$ es suficientemente grande, estamos en condiciones de aplicar el teorema central del límite para obtener que $\sum_{i=1}^{120} X_i \dot{\sim} N(120\lambda, 120\lambda) = N(1320, 1320)$, habiendo sustituido el valor desconocido de λ por la estimación obtenida en el apartado anterior.

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

$$p\left(\sum_{i=1}^{120} X_i > 1350\right) = p\left(Z > \frac{1350 - 1320}{\sqrt{1320}}\right) = p(Z > 0.8257) = \underline{0.2047},$$

con $Z = \frac{\sum_{i=1}^{120} X_i - 1320}{\sqrt{1320}} \dot{\sim} N(0, 1)$.

5. X_1, opinión de los jóvenes acerca de la emisión del anuncio A.
X_2, opinión de los jóvenes acerca de la emisión del anuncio B.

a) $H_0:$ X_1 y X_2 homogéneas
$H_1:$ no H_0.
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 de homogeneidad 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 α.

$$
\begin{aligned}
q \;\;=\;\; & \frac{\left(80-\frac{(150)(155)}{300}\right)^2}{\frac{(150)(155)}{300}}+\frac{\left(20-\frac{(150)(35)}{300}\right)^2}{\frac{(150)(35)}{300}}+\frac{\left(50-\frac{(150)(110)}{300}\right)^2}{\frac{(150)(110)}{300}}+ \\
& +\frac{\left(75-\frac{(150)(155)}{300}\right)^2}{\frac{(150)(155)}{300}}+\frac{\left(15-\frac{(150)(35)}{300}\right)^2}{\frac{(150)(35)}{300}}+\frac{\left(60-\frac{(150)(110)}{300}\right)^2}{\frac{(150)(110)}{300}}=1.7847,
\end{aligned}
$$

y como se tiene que el estadístico es menor que el cuantil, $q = 1.7847 < 4.605 = \chi^2_{2,0.1} = \chi^2_{(2-1)(3-1),0.1}$, la decisión es ACEPTAR H_0 para un nivel de significación $\alpha = 0.1$ y, por consiguiente, para cualquiera de los habituales.

b) $H_0 :\ \pi_{\cdot 1} = \pi^*_{\cdot 1} = 0.45,\ \pi_{\cdot 2} = \pi^*_{\cdot 2} = 0.1,\ \pi_{\cdot 3} = \pi^*_{\cdot 3} = 0.45$
   $H_1 :$ no H_0.
   El estadístico χ^2 de Pearson cumple que

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

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.

$$
q=\frac{(155-300\times 0.45)^2}{300\times 0.45}+\frac{(35-300\times 0.1)^2}{300\times 0.1}+\frac{(110-300\times 0.45)^2}{300\times 0.45}=8.4259,
$$

que es mayor que el cuantil $\chi^2_{2,0.05} = 5.991$, luego la decisión es RECHAZAR H_0 para un nivel de significación $\alpha = 0.05$. Ha de tomarse con prudencia esta decisión pues para un nivel de significación $\alpha = 0.01$ $\chi^2_{2,0.01} = 9.21$ y la decisión es justamente la contraria, ACEPTAR H_0.