OCW-UPM Estadística para Ingeniería Civil y Medioambiental

Autores: E. M. García del Toro, C. Hermoso, E. J. Huertas

LABORATORIO DE ESTADÍSTICA CON MATLAB

PRÁCTICA 6 - Inferencia estadística: estimación puntual y por intervalos de confianza. **PARTE B**

En esta práctica, seguimos trabajando los siguientes casos de *inferencia estadística*, de acuerdo con los contenidos del correspondiente tema de la asignatura. En esta ocasión hallaremos intervalos de confianza para la varianza poblacional de una distribución normal, e intervalos de confianza para la proporción.

En general, en el tema de *Estimación puntual y por intervalos de confianza*, los casos a estudio son los siguientes:

1) Estimación puntual:
   a) Estimación de parámetros poblacionales en la distribución binomial
   b) Estimación de parámetros poblacionales en la distribución normal
2) Estimación por intervalos de confianza:
   a) Intervalo de confianza para la media de una población de varianza conocida
   b) Intervalo de confianza para la media de una distribución normal de varianza desconocida
   c) Intervalo de confianza para la varianza poblacional de una distribución normal
   d) Intervalo de confianza para la proporción (distribución binomial)

Observación importante sobre notación: En lo que sigue, recuérdese que se denota por z_p, $t_{n;p}$, $\chi^2_{n;p}$, y $F_{n_1,n_2;p}$, respectivamente, el valor de la abcisa de una $N(0,1)$ normal estándar, t_n de Student, $\chi^2_{n;p}$ de Pearson, y $F_{n_1,n_2;p}$ de Fisher, que deja a su **derecha** -bajo la correspondiente función de densidad- un área de probabilidad p.

El cálculo de estas abcisas (también llamadas **valores críticos**) puede realizarse en MATLAB, respectivamente mediante los siguientes comandos:

```
norminv   %Inverse of the normal cumulative distribution function (cdf).
    X = norminv(P,MU,SIGMA) returns the inverse cdf for the normal
    distribution with mean MU and standard deviation SIGMA, evaluated at
    the values in P. Default values for MU and SIGMA are 0 and 1, respectively.

tinv      %Inverse of Student's T cumulative distribution function (cdf).
    X=tinv(P,V) returns the inverse of Student's T cdf with V degrees
    of freedom, at the values in P.

chi2inv   %Inverse of the chi-square cumulative distribution function (cdf).
    X = chi2inv(P,V)  returns the inverse of the chi-square cdf with V
    degrees of freedom at the values in P.

finv   %Inverse of the F cumulative distribution function.
    X=finv(P,V1,V2) returns the inverse of the F distribution
    function with V1 and V2 degrees of freedom, at the values in P.
```

También podremos hallar estos valores mediante la herramienta gráfica `disttool` que ya hemos visto en otras ocasiones, o en última instancia, si no disponemos de computador, mediante las tablas de las respectivas distribuciones.

Estimación por Intervalo de Confianza para la varianza σ^2 de una población normal

Considérese el mismo ejercicio de la práctica anterior, sobre el caudal de cierto químico que circula por la tubería de una fábrica:

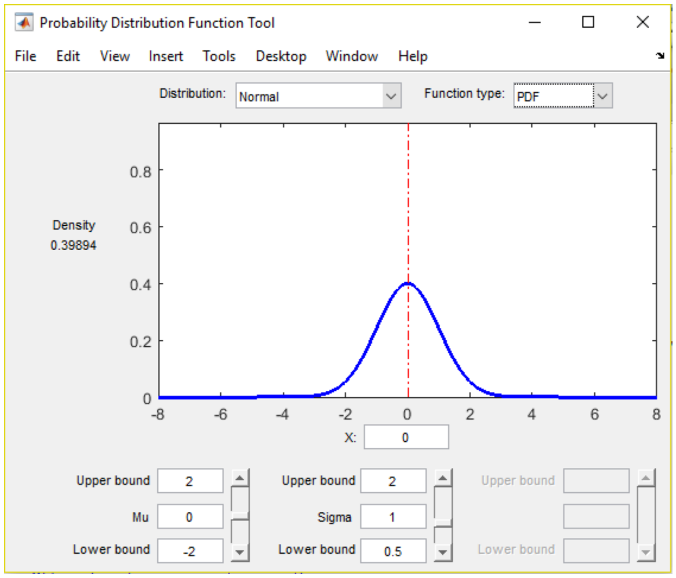

Figura 1: Ventana de la herramienta `disttool` en MATLAB

Se desea estimar el caudal medio μ (en metros cúbicos por segundo) de cierto producto químico que circula por la tubería de una fábrica. Se supone que dicho caudal sigue una distribución normal, de media μ y de varianza (poblacional) conocida $\sigma^2 = 144\ \left(m^3/s\right)^2$. Se llevan a cabo las siguientes 25 medidas (en m^3/s) de dicho caudal, en instantes elegidos al azar:

90.8, 81.5, 83.2, 79.7, 93.3, 65.0, 88.9, 90.9, 85.0, 72.5, 77.6, 101.4, 82.5, 93.6, 105.9, 71.1, 87.2, 75.0, 93.6, 61.8, 108.9, 88.7, 93.7, 80.6, 72.6

Ahora se va a tratar de obtener un Intervalo de confianza para la varianza poblacional de una distribución normal (lo que corresponde a **ejemplos del caso** $2c$))

a) Hallar un Intervalo de Confianza para la varianza poblacional σ^2 de la variable en estudio, con nivel de confianza del 99 %, considerando la media poblacional desconocida.

b) Hallar un Intervalo de Confianza para la varianza poblacional σ^2 de la variable en estudio, con nivel de confianza del 90 %, considerando la media poblacional desconocida.

Solución:
apartado a)

Si, como suele ocurrir en casos reales, no se conoce la media poblacional μ, la expresión para este Intervalo de Confianza es siempre:

$$I = \left[\frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}}\right], \tag{1}$$

donde $\chi^2_{n-1;\alpha/2}$ es el valor de la abcisa en la distribución Chi cuadrado de Pearson, de $n-1$ grados de libertad, que deja a su derecha un área de valor $\alpha/2$. Por otro lado, $\chi^2_{n-1;1-\alpha/2}$ será el valor de la abcisa en la distribución Chi cuadrado de Pearson, de $n-1$ grados de libertad, que deja a su derecha un área de valor $1-\alpha/2$.

Dado que en este caso la confianza es $1-\alpha = 0{,}99$, el valor de significación es $\alpha = 0{,}01$, por lo que $\alpha/2 = 0{,}005$. Como siempre, podemos calcular estos dos valores críticos en MATLAB, bien con la herramienta `disttool`, o bien escribiendo en la linea de comandos `chi2inv(1-0.005,24)` y `chi2inv(0.005,24)`, obteniendo:

$$\begin{aligned}
\chi^2_{n-1;0{,}005} &= \texttt{chi2inv}(1-0{,}005, 24) = 45{,}56, \\
\chi^2_{n-1;1-0{,}005} &= \texttt{chi2inv}(0{,}005, 24) = 9{,}886.
\end{aligned}$$

Reemplazando estos valores en (1), obtenemos:

$$I_{0{,}99} = \left[\frac{24 \cdot 139{,}04}{45{,}56}, \frac{24 \cdot 139{,}04}{9{,}886}\right] = [73{,}24\,,\, 337{,}5]\ \left(m^3/s\right)^2. \tag{2}$$

Nótese que, con un tamaño muestral tan pequeño como $n = 25$ muestras solamente, el intervalo de confianza al 99 % es tan grande que carece de utilidad práctica y resulta muy poco informativo. Además, debido a que la varianza se mide en unidades de la media **al cuadrado**, hay un efecto de distorsión en el tamaño del intervalo. Los resultados son más fáciles de asimilar y comprender si se expresa en términos de la desviación típica σ. Así, por ejemplo, el intervalo anterior (2), en términos de la desviación típica (que es simplemente la raíz cuadrada de la varianza) queda

$$\sigma^2 \in [73{,}24\,,\, 337{,}5]\ \left(m^3/s\right)^2 \to \sigma \in [\sqrt{73{,}24}\,,\, \sqrt{337{,}5}] = [8{,}56\,,\, 18{,}37]\ m^3/s$$

que, además de parecer de un tamaño más aceptable, es más lógico pues se expresa en las mismas unidades que la media.

apartado b)

Los cálculos aquí son prácticamente iguales que en el apartado anterior, solamente que tendremos que calcular otros valores críticos, ya que ha cambiado la confianza. Como en este apartado la confianza es $1 - \alpha = 0{,}90$, el valor de significación es $\alpha = 0{,}1$, por lo que $\alpha/2 = 0{,}05$. Así

$$\begin{aligned} \chi^2_{n-1;0{,}05} &= \texttt{chi2inv}(1 - 0{,}05, 24) = 36{,}415, \\ \chi^2_{n-1;1-0{,}05} &= \texttt{chi2inv}(0{,}05, 24) = 13{,}85. \end{aligned}$$

Sustituyendo estos valores en (1), obtenemos el nuevo intervalo de confianza:

$$I_{0{,}90} = \left[\frac{24 \cdot 139{,}04}{36{,}42}, \frac{24 \cdot 139{,}04}{13{,}85}\right] = [91{,}6\,,\, 240{,}9]\ \left(m^3/s\right)^2. \tag{3}$$

Obsérvese que, ni siquiera reduciendo el grado de confianza al 90 %, el intervalo correspondiente anterior (3) gana mucho en precisión. De nuevo, la razón es que la varianza se mide en unidades de la media **al cuadrado**, hay un efecto de distorsión en el tamaño del intervalo. Como en el caso anterior, es más fácil interpretar los resultados si se expresa (3) en términos de la desviación típica σ. Así,

$$\sigma^2 \in [91{,}6\,,\, 240{,}9]\ \left(m^3/s\right)^2 \to \sigma \in [\sqrt{91{,}6}\,,\, \sqrt{240{,}9}] = [9{,}57\,,\, 15{,}52]\ m^3/s$$

que, además de parecer más aceptable, es más lógico pues se expresa en las mismas unidades que la media.

Estimacion Puntual y por Intervalo de Confianza para la proporción poblacional

Un ayuntamiento desea conocer si en su ciudad se cumple la ley sobre la obligatoriedad del uso del casco reglamentario en motoristas. Para ello, encarga un estudio en el que se observa que, de 464 motoristas, 131 de ellos no llevan casco.

a) Hallar una estimación puntual de la proporción p de motoristas que circulan sin casco **(ejemplo del caso** $1a$**))**.

b) Hallar un intervalo de confianza al 90 % para la proporción poblacional p de motoristas que circulan sin casco **(ejemplo del caso** $2d$**))**. Indique qué condición se cumple para que el intervalo calculado sea válido.

c) Un dirigente de la oposición critica que la política de concienciación del uso del casco en la ciudad es pésima, ya que el 40 % de motoristas circulan actualmente sin este elemento imprescindible de seguridad ¿Se puede dar por cierta dicha afirmación al mismo nivel de confianza?

Solución:

apartado a) Según lo estudiado en las clases de teoría, el estimador más adecuado para la proporción poblacional desconocida p es la proporción muestral $\hat{p}$, donde

$$\hat{p} = \frac{\text{número de veces que aparece el suceso A}}{\text{tamaño muestral } n}.$$

Así, en nuestro caso

$$\hat{p} = \frac{131}{464} = 0{,}2823 \to 28{,}23\,\% \text{ de motoristas circulan sin casco.}$$

apartado b) Hay que hallar un intervalo de confianza al 90 % para la proporción poblacional, cuya expresión vimos en las clases teóricas que es siempre:

$$I = \left[\hat{p} \mp z_{\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\right] = \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\,,\,\hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\,,\right]. \tag{4}$$

El valor de la proporción muestral $\hat{p} = 0{,}2823$ lo hemos calculado en el apartado anterior. La confianza del 90 % indica que el nivel de significación es $\alpha = 0{,}1$ (ya que $90 = (1-\alpha)\cdot 100$). El valor crítico que se necesita en (4) es $z_{\alpha/2} = z_{0{,}1/2} = z_{0{,}05}$. Calculando este valor en MATLAB, por ejemplo mediante el comando `norminv`, obtenemos que $z_{0{,}05} =$ `norminv(1-0.05,0,1)` $= 1{,}645$. Así pues,

$$\begin{aligned}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}} &= \sqrt{\frac{0{,}28233\,(1-0{,}28233)}{464}} = \sqrt{\frac{0{,}28233\cdot 0{,}717\,67}{464}} \\ &= \sqrt{4{,}366\,8\times 10^{-4}} = 0{,}020897.\end{aligned}$$

Reemplazando los cálculos en (4) tenemos

$$\begin{aligned} I_{0{,}90} &= \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\,;\,\hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\right] \\ &= [0{,}28233 - 1{,}645\cdot 0{,}020897\,;\,0{,}28233 + 1{,}645\cdot 0{,}020897] \\ &= [0{,}248\,;\,0{,}317]\,. \end{aligned}$$

Para que este intervalo de confianza sea válido, el tamaño muestral debe ser mayor que 30. En este caso es 464, por lo que esta condición se cumple.

apartado c) Como el valor del 40 % (que corresponde a una proporción del 0,4) **no** se encuentra en el interior del intervalo de confianza calculado en el apartado a), **no puede darse por cierta** la afirmación a un nivel de confianza del 90 %.

Trabajo para el alumno: Resolver el siguiente ejercicio:

En una encuesta para estudiar el grado de preocupación de la población por el medio ambiente, se ha preguntado a 460 personas de las que 130 han contestado que reciclan los envases de plástico que utilizan.

a) Hallar una estimación puntual de la proporción p de personas que reciclan.

b) Hallar un intervalo de confianza al 90 % para la proporción poblacional de personas que reciclan. Indique qué condición se cumple para que el intervalo calculado sea válido.

c) En las noticias se publica un titular que afirma que el 35 % de personas de la población reciclan los envases de plástico. ¿Se puede dar por cierta dicha afirmación al mismo nivel de confianza?

a) Hay que hallar una estimación puntual del parámetro poblacional p. Según lo estudiado en las clases de teoría, el estimador más adecuado para la proporción poblacional desconocida p es la proporción muestral $\hat{p}$, donde

$$\hat{p} = \frac{\text{número de veces que aparece el suceso A}}{\text{tamaño muestral } n}.$$

Así, en nuestro caso

$$\hat{p} = \frac{130}{460} = 0{,}2826 \to 28{,}26\,\% \text{ de personas que reciclan.}$$

b) Hay que hallar un intervalo de confianza al 90 % para la proporción poblacional, cuya expresión vimos en las clases teóricas que es siempre:

$$I = \left[\hat{p} \mp z_{\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\right]. \tag{5}$$

La confianza del 90 % indica que el nivel de significación es $\alpha = 0{,}1$ (ya que $90 = (1-\alpha)\cdot 100$). El valor crítico que se necesita en (5) es $z_{\alpha/2} = z_{0,1/2} = z_{0,05}$. Mirando este valor en la tabla de la normal estándar, obtenemos que $z_{0,05} = 1{,}645$. El valor de la raíz es:

$$\begin{aligned}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}} &= \sqrt{\frac{0{,}28261\,(1-0{,}28261)}{460}} \\ &= \sqrt{\frac{0{,}28261\cdot 0{,}71739}{464}} = \sqrt{4{,}3694\times 10^{-4}} = 0{,}020903.\end{aligned}$$

Reemplazando los cálculos en (5) tenemos

$$\begin{aligned} I_{0,90} &= \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\,;\,\hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\right] \\ &= [0{,}28261 - 1{,}645\cdot 0{,}020903\,;\,0{,}28261 + 1{,}645\cdot 0{,}020903] \\ &= [0{,}28261 - 0{,}034385\,;\,0{,}28261 + 0{,}034385] \\ &= [0{,}24822\,;\,0{,}31700]\,. \end{aligned}$$

Para que este intervalo de confianza sea válido, el tamaño muestral debe ser mayor que 30. En este caso tenemos $n = 460 > 30$, por lo que este valor se cumple.

c) Como el valor del 35 % (que corresponde a una proporción del 0,35) **no** se encuentra en el interior del intervalo de confianza calculado en el apartado a), **no puede darse por cierta** la afirmación a un nivel de confianza del 90 %.