Métodos Estadísticos
Tarea 2

Rubén Pérez Palacios Lic. Computación Matemática
Profesor: Dr. Rogelio Ramos Quiroga

17 de febrero de 2022

En el área de microbiología cuantitativa es de interés determinar la densidad de microorganismos presentes en un determinado medio. Sea θ el número medio de microorganismos por unidad de volumen en cierto medio líquido. Queremos estimar θ.

Pensemos en una unidad de volumen (digamos 1 cm^3), y supongamos que el número de microorganismos, Y, por unidad de volumen sigue una ley Poisson,

$$P(Y = y) = \frac{e^{-\theta}\theta^y}{y!}, \qquad y = 0, 1, 2, \cdots,$$

donde θ es el número medio por unidad de volumen. La forma obvia es tomar una muestra, digamos 10 cm^3, del líquido original y contar el número de microorganismos, N, que hay en la muestra y entonces $\widehat{\theta} = N/10$ y listo $\cdots$ un pequeño problema es que, en general, esto no se puede hacer en forma directa (en algunos casos, la tecnología todavía no puede hacer ese conteo), entonces hay que recurrir a métodos indirectos.

Entonces la idea es contar el número de platos, de los n, en los que no hay actividad biológica (platos inertes), sea X el número de platos inertes, entonces X sigue una distribución Binomial

$$P(X = x) = \binom{n}{x} p^x(1-p)^{n-x}, \qquad x = 0, 1, 2, \cdots, n,$$

de aquí que el problema se reduce a estimar p (fácil: $\widehat{p} = X/n$), pues si tenemos $\widehat{p}$ entonces como $p = e^{-\theta}$, tendremos que $\widehat{\theta} = -\log(\widehat{p})$ y listo $\cdots$

En principio la idea es buena pero hay un pero. Si, por ejemplo, tenemos 6 platos de Petri, probablemente vamos a observar:

actividad en plato 1 = si,	actividad en plato 2 = si,
actividad en plato 3 = si,	actividad en plato 4 = si,
actividad en plato 5 = si,	actividad en plato 6 = si

pues es muy probable que en cada unidad de volumen se vaya al menos un bicho, lo cual es suficiente para que, al ponerlo con nutientes, observemos actividad biológica.

En general, es difícil estimar la probabilidad de éxito (plato inerte) si sólo observamos fracasos!. En fin, para sacarle la vuelta a este problema tenemos dos caminos: Aumentar n o tomar muestras diluídas.

Consideremos la segunda opción. Aquí aumentamos la probabilidad de tener una muestra inerte simplmente diluyendo la muestra original, por ejemplo, tomamos 1 cm^3 del líquido original y lo mezclamos con 1 cm^3 de un líquido limpio y ponemos en un plato la mitad de esa mezcla la cual tendrá una densidad media igual a $\theta/2$. Podemos continuar de esta forma tomando muestras más y más diluídas, en realidad no se tienen que diluír las muestras a la mitad, sea a el factor de disolución ($a = 2$ en el caso anterior); tendremos entonces (después de k disoluciones) muestras con densidades (número de microorganismos por unidad de volumen):

$$\frac{\theta}{a^0}, \quad \frac{\theta}{a^1}, \quad \frac{\theta}{a^2}, \quad \cdots \quad \frac{\theta}{a^k}.$$

Supongamos que al nivel 0 de disolución observamos x_0 platos inertes (de un total de n_0 platos); al nivel 1 de disolución tenemos x_1 platos inertes (de un total de n_1) y así sucesivamente. En cada paso, el número de platos inertes tendrá una distribución Binomial con parámetros n_i y $p_i = \exp(-\theta/a^i)$.

Entonces, la verosimilitud para nuestros datos es

$$L(\theta) = \prod_{i=0}^{k} f(x_i|\theta) = \prod_{i=0}^{k} \begin{pmatrix} n_i \\ x_i \end{pmatrix} p_i^x (1 - p_i)^{n_i - x_i}, \text{ donde } p_i = \exp\left(\frac{-\theta}{a^i}\right).$$

En un experimento de este tipo se usó $a = 2$, $k = 9$ y, a cada nivel de disolución se usó $n_i = 5$. Los datos (número de platos inertes) observados son:

$$\{x_i\} = \{0, 0, 0, 0, 1, 2, 3, 3, 5, 5\}$$

1. Encuentre el estimador de máxima verosimilitud para θ usando el método Newton-Raphson.

   El estimador de máxima verosimilitud para θ es:

   $$\hat{\theta} = 30.64979,$$

   y es máximo puesto que

   $$-I_O = -0.0122.$$

2. Encontrar el intervalo de verosimilitud del 25 % para θ.

   El intervalo del 25 % para θ es:

   $$IV_{0.25}(\theta) = \{\theta : R(\theta; x) \geq 0.10\} = [12.940345430114384, 48.35923539576445].$$

3. La varianza del estimador de máxima verosimilitud (resultado asintótico) está dada por $I^{-1}(\theta)/k$ (equivalentemente $[kI(\theta)]^{-1}$, donde

   $$I(\theta) = E\left(-\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2}\right),$$

   note que esta expresión supone observaciones independientes e idénticamente distribuidas (el cual no es nuestro caso: tenemos k diferentes binomiales en la verosimilitud). Una forma de aproximar la varianza del estimador máximo verosímil es tomando el inverso de la "matriz de información observada":

   $$I_o = -\frac{\partial^2 \log L(\widehat{\theta})}{\partial \theta^2} = -\sum_{i=1}^{k} \frac{\partial^2}{\partial \theta^2} \log f(x_i|\theta)|_{\theta=\widehat{\theta}}.$$

   Así, la varianza se puede aproximar como $V(\widehat{\theta}) = I_o^{-1}$. Calcule esta varianza.

   La varianza es

   $$V(\theta) = 81.6390154101694.$$

4. Usando el resultado obtenido en el inciso anterior, obtenga un intervalo del 95 % de confianza para θ:

$$\widehat{\theta} \pm z_{\alpha/2}\sqrt{I_o^{-1}}.$$

Compare y comente este intervalo con el del inciso 2.

Un intervalo del 95 % de confianza para θ es:

$$[12.940345430114384, 48.35923539576445] .$$

5. El experimento completo usó 9 disoluciones (10 contando la primera en donde no hay disolución), en cada etapa de usaron 5 platos de Petri, dando un total de 50 platos en total. Una pregunta natural es: ¿Valió la pena el trabajo de hacer las disoluciones?, ¿no hubiera sido mejor simplemente preparar 50 platos de Petri con muestras sin diluír y en base a ellos estimar p? ... Explique cómo respondería a esta pregunta usando simulación (no hay que hacer la simulación, simplemente explicar como le podría hacer).

Dada una simulación calcularía el máximo estimador de verosimilitud de *theta* cuando las muestras no se disimulan, llamemoslo $\hat{\theta}_1$ y $\hat{\theta}$ de la misma simulación, después calcularía sus respectivos intervalos de confianza. Además generaría varias simulaciones como la obtenida en el problema $(n = 5, k = 9, \cdots)$ y varias con solo una disolución y aquel intervalo de confianza que tenga un valor mayor o más cercano a 95 % sería el criterio de desempate sería el tenga un longitud menor.