NOTAS ACERCA DE LAS FÓRMULAS DE SIGNIFICANCIA ESTADÍSTICA

1. Brevísimo contexto estadístico

Si tomamos como referencia los papers de K. Cranmer, G. Cowan, E. Gross y O. Vitells, así como también libros como el del mismo Cowan, tenemos que, entro otros, los test estadísticos más utilizados para cuantificar descubrimiento y para imponer límites superiores son los llamados q_0 y q_μ y vienen definidos por:

$$q_0 = \begin{cases} -2\ln(L(0)/L(\hat{\mu})) & \hat{\mu} \geq 0, \\ 0 & \hat{\mu} < 0, \end{cases} \tag{1}$$

$$q_\mu = \begin{cases} -2\ln(L(\mu)/L(\hat{\mu})) & \hat{\mu} \leq \mu, \\ 0 & \hat{\mu} > \mu, \end{cases} \tag{2}$$

donde L es la función de likelihood y $\hat{\mu}$ hace referencia al valor del parámetro μ que maximiza dicha función. A partir de estos test es posible definir los correspondientes p-values como sigue

$$p_0 = \int_{q_{0obs}}^{\infty} f(q_0|0)dq_0, \tag{3}$$

donde $f(q_0|0)$ es la pdf para q_0 bajo la hipótesis de solo fondo, y

$$p_\mu = \int_{q_{\mu obs}}^{\infty} f(q_\mu|\mu)dq_\mu, \tag{4}$$

con $f(q_\mu|\mu)$ la pdf de q_μ suponiendo la hipótesis μ. A partir de los valores p es posible obtener un significancia a partir de la relación

$$Z = \Phi^{-1}(1-p), \tag{5}$$

donde Φ^{-1} es el cuantil o inversa de la distribución cumulativa de la Gaussiana estándar. De esta forma una significancia de 5σ, o sea $Z = 5$, se corresponde con $p = 2.87 \times 10^{-7}$ y se toma como un nivel apropiado para establecer descubrimiento, mientras que $Z = 1.64$ corresponde a $p = 0.05$ y daría lugar a excluir una determinada hipótesis de señal con un 95% de nivel de confianza.

Ahora bien, a la hora de caracterizar la sensibilidad de un experimento, la cantidad relevante deja de ser la significancia obtenida para un conjunto de datos específico. En su lugar lo que se busca es la significancia esperada con la que uno sería capaz de rechazar, por ejemplo, distintos valores de μ (hipótesis de señal). Más concretamente, en el caso de descubrimiento es relevante conocer la mediana con la que, bajo la hipótesis de una señal nominal ($\mu = 1$), uno podría rechazar la hipótesis de solo fondo ($\mu = 0$). En el caso de exclusión se busca determinar la mediana, bajo a hiopótesis de solo fondo, con la cual es posible rechazar un valor no nulo de μ. Para obtener las correspondientes medianas, se utilizan las siguientes expresiones:

$$\text{med}[Z_0|\mu'] = \sqrt{q_{0,A}}, \tag{6}$$

$$\text{med}[Z_\mu|0] = \sqrt{q_{\mu,A}}, \tag{7}$$

donde el subíndice A hace referencia a que los test q deben evaluarse en el conjunto de datos conocido como Asimov, el cual se define como aquel tal que al ser utilizado para evaluar los estimadores de todos los parámetros, uno obtiene los verdaderos valores de los parámetros. Si se considera un experimento de conteo en bines, por ejemplo, y suponemos distribuciones Poissoneanas, entonces fijando una cierto valor de μ, sea μ', el número esperado de eventos en cada bin sería $\nu_i = \mu' s_i + b_i$

y entonces el conjunto de datos, es decir, el conjunto de número de eventos en cada bin n_i que nos devolvería el parámetro μ' al maximizar el likelihood viene dado por

$$n_{i,A} = E[n_i] = \nu_i = \mu' s_i + b_i, \tag{8}$$

lo cual define el conjunto de datos Asimov. Este es el conjunto que debe usarse para evaluar los test estadísticos en las ecs. (6)-(7).

2. Fórmula de significancia para evidencia y/o descubrimiento

Deduzcamos ahora la fórmula de significancia para cuantificar sensibilidad de evidencia o descubrimiento. Suponemos un experimento donde el número de eventos de fondo es conocido, sea b, y el número de eventos de señal se parametriza como μs. La función de likelihood en este caso es

$$L(\mu) = \frac{(\mu s + b)^n}{n!} e^{-(\mu s + b)}. \tag{9}$$

Siguiendo la ec. (1) debemos calcular la cantidad $-2\ln(L(0)/L(\hat{\mu}))$. La maximización del likelihood da en este caso $\hat{\mu} = (n-b)/s$. Entonces:

$$-2\ln\left(\frac{L(0)}{L(\hat{\mu})}\right) = 2\left(n\ln\frac{n}{b} + b - n\right),$$

de manera que evaluando en el set de Asimov para el valor nominal $\mu' = 1$, es decir, reemplazando n por $s + b$ (ver ec. (8)), se tiene

$$\text{med}[Z_0|1] = \sqrt{q_{0,A}} = \sqrt{2((s+b)\ln(1+s/b) - s)}. \tag{10}$$

Nótese que esta es la expresión que venimos utilizando para cuantificar evidencia o descubrimiento (sin tener en cuenta posibles incertezas sistemáticas en el fondo). Estudiemos ahora el límite de la ec. (10) cuando $s << b$. En tal caso, desarrollando el logaritmo a segundo orden en s/b, es decir, usando

$$\ln\left(1 + \frac{s}{b}\right) \approx \frac{s}{b} - \frac{s^2}{2b^2},$$

obtenemos

$$\text{med}[Z_0|1] \approx \frac{s}{\sqrt{b}}, \tag{11}$$

la cual constituye la expresión vastamente utilizada para computar significancias de manera naif.

3. Fórmula de significancia para límites de exclusión

Efectuemos ahora un cálculo similar para el caso de límites de exclusión. En este caso, según la ec. (2), debemos calcular $-2\ln(L(\mu)/L(\hat{\mu}))$. Por tanto,

$$-2\ln\left(\frac{L(\mu)}{L(\hat{\mu})}\right) = 2(n\ln n - n - n\ln(\mu s + b) + \mu s + b).$$

Según la ec. (7) debemos ahora utilizar el set de Asimov pero para $\mu' = 0$, es decir, debemos reemplazar en la expresión anterior n por b (ver ec. (8)) para así obtener

$$\text{med}[Z_\mu|0] = \sqrt{q_{\mu,A}} = \sqrt{2\left(b\ln\left(\frac{b}{s+b}\right) + s\right)}, \tag{12}$$

donde en un abuso de notación hemos llamado s a lo que antes era μs, dado que esta última es la cantidad de eventos de señal. Nótese que es esta la expresión que uno debiese utilizar para calcular límites de exclusión esperados a una cierta luminosidad y que la misma difiere de la correspondiente a evidencia o descubrimiento (ec. (10)). El hecho de que a veces se considere conjuntamente ambas nociones tiene que ver nuevamente con el límite $s << b$. Si ahora desarrollamos el logaritmo se tiene

$$\text{med}[Z_\mu|0] \approx \sqrt{2\left(s - b\left(\frac{s}{b} - \frac{s^2}{2b^2}\right)\right)} = \frac{s}{\sqrt{b}}, \tag{13}$$

es decir, obtenemos exactamente la misma expresión de la ec. (11). Ahora bien, esta expresión difiere de otra también ampliamente utilizada, a saber $s/\sqrt{s+b}$. Mostraremos ahora que dicha expresión proviene del límite $s + b >> s$. Para ello utilizaremos el siguiente desarrollo para el logaritmo que aparece en la ec. (12)

$$\ln\left(\frac{s+b}{b}\right) = -\ln(1 - \frac{s}{s+b}) \approx \frac{s}{s+b} - \frac{s^2}{2b^2}.$$

Reemplazando obtenemos para la mediana

$$\sqrt{2\left(s - \frac{bs}{s+b} + \frac{bs^2}{2(s+b)^2}\right)} = \sqrt{2\left(\frac{s^2}{s+b} - \frac{b}{2}\frac{s^2}{(s+b)^2}\right)} = \sqrt{2\left(\frac{s^2}{(s+b)}\left(1 - \frac{b}{2(s+b)}\right)\right)},$$

pero la expresión dentro del último paréntesis puede reescribirse como

$$1 - \frac{b}{2(s+b)} = 1 - \frac{s+b-s}{2(s+b)} = \frac{1}{2} - \frac{s}{2(s+b)}$$

de manera que la mediana queda

$$\sqrt{\frac{s^2}{(s+b)} - \frac{s^3}{(s+b)^2}}$$

, lo cual al retener solo el término dominante nos da finalmente

$$\text{med}[Z_\mu|0] \approx \frac{s}{\sqrt{s+b}}, \tag{14}$$

válida en el límite $s + b >> s$. Por supuesto, si además se toma el límite $s << b$ en la ec. (14), recuperamos la expresión ya dada en la ec. (13).