TRATAMIENTO DE DATOS DE UNA MAGNITUD QUE PRESENTA DISTRIBUCIÓN NORMAL TOMADOS CON ERROR INSTRUMENTAL NO DESPRECIABLE

J. BLOSTEIN* P. Do CAMPO* C. MORENO**

INSTITUTO DE FÍSICA DEL PLASMA, CONICET,
FACULTAD DE CIENCIAS EXACTAS Y NATURALES, DEPARTAMENTO DE FÍSICA,
UNIVERSIDAD DE BUENOS AIRES, CIUDAD UNIVERSITARIA PAB 1,
1428 BUENOS AIRES.

Se desarrolla un formalismo que permite estimar la esperanza y la varianza de variables aleatorias obtenidas con incerteza de truncamiento. El método se basa en la función de distribución de la variable truncada y en el criterio de máxima verosimilitud. Se demuestra también que cuando el truncamiento es relevante, el promedio muestral no es un estimador consistente de la esperanza de la variable medida.

Introducción

No siempre resulta inmediato determinar la incerteza de una medición directa. Una de las situaciones más delicadas se presenta cuando la incerteza debida al truncamiento del sistema de medición, $\Delta/2$, es comparable a las variaciones propias, σ, de la magnitud medida. En estos casos la dispersión muestral, inevitablemente afectada por Δ, no es un estimador de σ. Para encarar el problema de la estimación del valor de la magnitud medida y de su dispersión, es conveniente considerar las siguientes situaciones.

A) Si se midiese una magnitud que no presenta dispersión estadística con un instrumento ideal, esto es, que no introduce errores, se podría informar exactamente el valor de la magnitud, y como ese valor sería exacto, y se sabría que lo es, se le asignaría error nulo.

B) Si se mide una magnitud que no presenta dispersión estadística con un instrumento que introduce error de truncamiento, no se puede informar exactamente el valor de la magnitud, sólo se puede afirmar con seguridad que la magnitud se encuentra en un determinado intervalo. En este caso, a pesar de que se incremente el número de mediciones, el valor informado por el instrumento siempre será el mismo. Para afirmar lo anterior se asumió que tanto la magnitud como el método de medición no presentan dispersión estadística. En el caso de utilizar un instrumento con divisiones de escala equiespaciadas, asumimos que el método de medición da como parte de su resultado la división de la escala que se encuentra más próxima a la magnitud medida. En otras palabras el método consiste en un redondeo. La otra parte del resultado es el error asociado a esa medición: el error asignado habitualmente es la mitad de la distancia entre las divisiones de la escala. Se destaca que este error no proviene de calcular la varianza de ninguna función de densidad de probabilidad (f.d.p.).

C) Al medir con error de truncamiento una magnitud fija, empleando un sistema de medición que presenta algún tipo de dispersión estadística, no es posible informar exactamente el valor de la magnitud. Se puede informar como resultado de la medición los momentos de primer y segundo orden de los datos obtenidos. En el caso de suponer que el método presenta distribución normal de parámetros μ y σ, $N(\mu,\sigma)$, se verá que el momento de primer orden no es un estimador consistente de μ; esto

* Dpto. de Física FCEyN, UBA.
** INFIP-CONICET, Dpto. de Física FCEyN UBA.

es, aún en el límite de infinitas mediciones el promedio muestral puede diferir de μ.

D) Por último, cuando se mide una magnitud que presenta dispersión estadística de algún tipo con un instrumento que tiene error de truncamiento, tampoco es posible informar exactamente algún valor particular de la magnitud en cuestión porque está descripta por una f.d.p. Justamente interesa conocer lo mejor posible los parámetros de esa f.d.p., que están determinados ya sea por el método; por variaciones propias de la magnitud medida; o por ambos factores. El tratamiento de los datos experimentales deberá dar por resultado un valor y una incerteza para cada parámetro de la f.d.p.

Este trabajo consiste en desarrollar un formalismo aplicable a las dos últimas situaciones expuestas, es decir, que trate datos de una magnitud que presenta dispersión normal, tomados con un instrumento que introduce error de truncamiento.

Desarrollo

Tanto en el caso en que las fluctuaciones de las lecturas efectuadas con un instrumento con error de truncamiento sean debidas a variaciones de la magnitud en cuestión como al sistema de medición, el tratamiento matemático propuesto será el mismo.

El desarrollo matemático parte de suponer una variable aleatoria X distribuida normalmente con parámetros μ y σ que se desean determinar. El instrumento de medición se modelará de la siguiente manera: tiene divisiones equiespaciadas a intervalos de ancho Δ. Cuando se realiza una medición, la variable aleatoria X adopta un valor x. Si x pertenece al rango del instrumento se obtiene una lectura. Esta lectura será aquella división que se encuentre más cercana a x (Fig. 1).

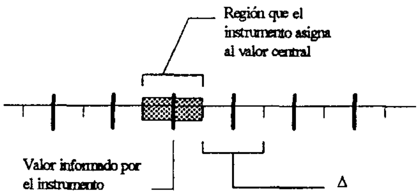

Fig. 1: Esquema del instrumento de medición.

La escala del instrumento tendrá una posición relativa fija respecto de μ. Tomando las lecturas del instrumento como eventos de interés, la probabilidad de obtener una lectura x cualquiera, es la probabilidad de que X adopte cualquier valor dentro del intervalo $[x - \Delta/2 , x + \Delta/2)$:

$$g(x) = \int_{x-\Delta/2}^{x+\Delta/2} \frac{e^{\frac{-(t-\mu)^2}{2.\sigma^2}}}{\sqrt{2.\pi}.\sigma} dt$$

La función g(x) depende paramétricamente de μ, σ y Δ. Si se tiene entonces un instrumento cuya resolución es Δ y si se admite que $i.\Delta$ es una lectura posible ($i \in Z$, donde i es el indicador del intervalo), entonces la probabilidad de obtener la lectura $i.\Delta$ en una medición será $g(i\Delta)$. Esto significa que la función g(x) al ser evaluada en los valores correspondientes a las lecturas del instrumento representa un histograma (teórico, de variable discreta) que indica la frecuencia relativa con la que debería obtenerse cada lectura. Puede verificarse que $g(i\Delta) \geq 0 \; \forall i$ y que

$$\sum_{i=-\infty}^{\infty} g(i\Delta) = 1 \quad \forall \Delta \in R_{>0}$$

con lo cual queda garantizado que $g(i\Delta)$ es una función de probabilidad discreta[1].

Obteniendo la curva teórica que mejor se ajuste[†] al histograma experimental es posible estimar μ y σ. Dado un instrumento, la biyectividad entre cada $N(\mu,\sigma)$ y cada $g(i.\Delta)$ se

† Entendemos por mejor ajuste aquel que surge de aplicar el criterio de ML u otro equivalente que se especifique.

puede demostrar matemáticamente en base a que μ y σ son independientes (un corrimiento en μ no puede compensarse modificando σ).

Para encontrar la función g que mejor se ajusta al histograma experimental se aplicará el criterio de máxima verosimilitud[2] (ML). Para cada lectura, pueden ocurrir sólo dos eventos: se leyó o no se leyó $i.\Delta$. La probabilidad de obtener $i.\Delta$ es $g(i.\Delta)$.

El número de formas distintas de que, una vez realizadas N lecturas, n_i correspondan a la i-sima división está dada por

$$\frac{N!}{\prod_{i=-\infty}^{+\infty} n_i!}$$

y la función de verosimilitud resulta

$$L = N! \prod_{i=-\infty}^{\infty} \frac{g(i.\Delta)^{n_i}}{n_i!}$$

L es la probabilidad de obtener un histograma experimental, en el cual no interesa el orden en que se realizaron las distintas lecturas, sino los totales por clase n_i.

Maximizando L es posible obtener los parámetros de $g(i.\Delta)$. Aplicando logaritmos, la expresión a maximizar resulta

$$\ln L = C + \sum_{i=-\infty}^{\infty} n_i \ln g(i.\Delta)$$

donde

$$C = \ln \frac{N!}{\prod_{i=-\infty}^{\infty} n_i!}$$

no depende de las incógnitas μ y σ.

Finalmente se define la función objetivo

$$S \equiv -\ln L + C = -\sum_{i=-\infty}^{\infty} n_i \ln g(i.\Delta) \qquad (1)$$

que depende de los parámetros de g: μ, σ y Δ. Una vez realizadas N mediciones (y en consecuencia, conocidas n_i), se toman como estimadores de μ y σ aquellos que minimizan S, (esto es, maximizan L) asumiendo Δ conocida a partir de las características del instrumento. Si bien estos estimadores pueden resultar sesgados, en el límite $N \rightarrow \infty$ tienden a los parámetros μ y σ de la gaussiana original puesto que son estimadores de ML[2]. La minimización de S no tiene solución analítica pero es de fácil tratamiento numérico. Algoritmos de uso difundido, tales como los basados en técnicas Quasi-Newton o Nelder-Meade[3] se desempeñan satisfactoriamente.

El promedio muestral

Puede verificarse que la esperanza matemática de la variable truncada, E, se relaciona con μ mediante

$$\mu = E + (\Delta / 2)\, f(\delta,\sigma) \qquad (2)$$

siendo

$$f(\delta,\sigma) = 2 \sum_{i=-\infty}^{+\infty} (i_o - i + \frac{\delta}{\Delta})\, g(i\,\Delta)$$

donde δ es la diferencia entre μ y el valor de lectura, $i_o\,\Delta$, más próximo a μ (esto es, $\mu = i_o\,\Delta + \delta$).

La expresión (2) establece que en presencia de errores de truncamiento ($\Delta \neq 0$) el promedio muestral, que converge a E a medida que se incrementa el número de datos[2], es un estimador consistente de μ si y sólo si $f(\delta,\sigma)=0$.

El gráfico de f puede verse en la Fig. 2. No se ha graficado f en el dominio $\delta \in (-0.5,0)$ pues es impar como función de δ. Se observa que cuando $\delta \sim 0.5$, $f \rightarrow 1$ a medida que σ/Δ disminuye, con lo cual, E dista de μ en $\Delta/2$. Por lo tanto, si $\sigma \ll \Delta$, sólo puede afirmarse que $\mu = E \pm \Delta/2$, dado que δ es desconocido.

Puede verificarse que $f \rightarrow 0$ cuando $\sigma/\Delta \gg 1$, lo cual indica que cuando la incerteza de origen estadístico es dominante, el error de

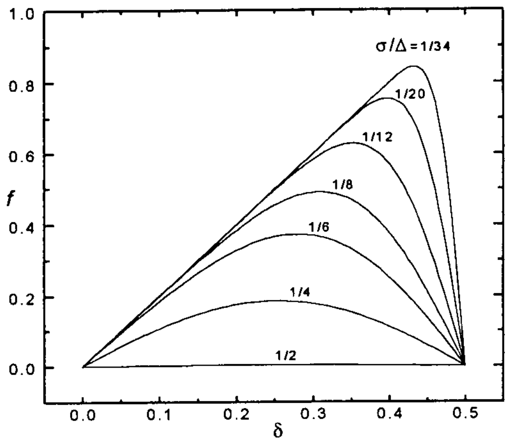

Fig 2: f en función de δ para varios valores de σ/Δ.

truncamiento no afecta la estimación de μ, como cabe esperar.

De la Fig. 2, se ve que los máximos de f aumentan a medida que σ disminuye (a Δ fijo). De esta forma si se dispone de una cota inferior para σ, σ_{min}, es posible afirmar que μ pertenece al intervalo centrado en E de amplitud $\Delta/2\ \mathrm{Max}[f(\delta,\sigma_{min})]$. Un valor para σ_{min} puede estimarse a partir de la varianza muestral, o bien de analizar la convergencia de los estimadores de σ conforme aumenta el número de datos.

Conclusiones

La determinación de μ y σ a partir de mediciones afectadas por incerteza de truncamiento puede efectuarse minimizando la expresión (1), que es válida en el caso general. La convergencia de los valores que de ella se obtienen, a medida que se incrementa el número de datos, está garantizada por ser de ML.

En principio es posible aproximarse tanto como se quiera a los parámetros μ y σ, aumentando el número de mediciones, independientemente de la incerteza de truncamiento del instrumento. Ello se debe a que además de las frecuencias relativas de las lecturas, se cuenta con el conocimiento de la f.d.p. que las describen incluyendo el truncamiento.

El límite práctico se da cuando el truncamiento es de tal magnitud, comparado con σ, que todas las mediciones arrojan el mismo resultado. En ese caso, sólo se puede afirmar que $\mu = E \pm \Delta/2$.

A partir de los valores de $f(\delta,\sigma)$ ilustrados en la Fig. 2 puede verse que si $\Delta \leq 2\ \sigma$ el efecto del truncamiento es irrelevante, y en consecuencia el promedio y la dispersión muestrales pueden tomarse como estimadores de μ y σ, respectivamente. En otro caso es conveniente estimar μ y σ minimizando (1).

Agradecimientos: Uno de nosotros (CM) agradece a R. Boquet por las discusiones mantenidas sobre temas de este artículo. Este trabajo fue subsidiado parcialmente por el PID UBA EX043 y por el CONICET.

Referencias

1- Feller W., *An introduction to probability theory and its applications*, Vol 1, Cap 1, 3° Ed. New York, Wiley, (1968).

2- Meyer P. L., *Probabilidad y aplicaciones estadísticas*, Cap 14, México, Fondo Educativo Interamericano, (1973).

3- Press W., Teukolsky S., Vetterling W., Flannery B., *Numerical recipes: The art of scientific computing*, Cap 10, 2° Ed. Cambridge (1992).