DISTRIBUCIÓN DE LA MEDIA Y EL TEOREMA DEL LÍMITE CENTRAL

Wolfgang A. Schmid
Centro Nacional de Metrología
Tel.: (442) 211 0554, e-mail: wschmid@cenam.mx

Resumen: De acuerdo al Teorema del Límite Central, la distribución de la media $\overline{X}$ de una serie de mediciones repetidas (X_1, X_2, … X_n) se aproxima a una distribución normal, independientemente de la distribución de los datos originales x_i. Este resultado importante para la estimación de la incertidumbre en mediciones se prueba con datos aleatorios provenientes de diferentes distribuciones. En particular, se analiza el caso de datos medidos con un instrumento con resolución "burda" y la interacción entre la dispersión de los datos cuantificada por su desviación estándar y la resolución del instrumento en la incertidumbre combinada. Se observa un incremento de la dispersión de los datos como efecto del redondeo, que con los conceptos de metrología se interpreta como una combinación de las incertidumbres por repetibilidad y resolución.

1. INTRODUCCIÓN

El mejor estimado de una magnitud X con un error asociado que varía de forma aleatoria generalmente es la media $\overline{X}$ de un número n de mediciones independientes X_i realizadas bajo las mismas condiciones:

$$\overline{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i \qquad (1)$$

La incertidumbre u de X se estima mediante la desviación estándar experimental de la media:

$$u_A(X) = s(\overline{X}) = \frac{s(X_i)}{\sqrt{n}} = \sqrt{\frac{1}{n \cdot (n-1)} \cdot \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} \qquad (2)$$

donde $s(X_i)$ es la desviación estándar experimental de las mediciones individuales. El índice A indica que se trata de un "método tipo A" para estimar la incertidumbre.

Para determinar el nivel de confianza relacionado con la incertidumbre de una magnitud, es necesario conocer su distribución. Del Teorema del Límite Central resulta que la distribución de la media $\overline{X}$ de una serie de mediciones repetidas e independientes (X_1, X_2,…X_n) se aproxima a una distribución normal, independientemente de la distribución de los X_i.

Para resaltar esto, véase por ejemplo [1], G.2.1.:

"Si $Y = c_1 X_1 + c_2 X_2 + ... c_N X_N = \sum_{i=1}^{N} c_i X_i$ y todas las X_i se caracterizan mediante distribuciones normales, entonces la distribución resultante de la convolución Y también será normal. Sin embargo, aún si las distribuciones de X_i no son normales, la distribución de Y frecuentemente se puede aproximar mediante una distribución normal debido al Teorema del Límite Central. Este teorema establece que la distribución de Y será aproximadamente normal con esperanza $E(Y) = \sum_{i=1}^{N} c_i E(X_i)$ y varianza $V(Y) = \sum_{i=1}^{N} c_i^2 \cdot V(X_i)$ donde $E(X_i)$ es la esperanza de X_i y $V(X_i)$ es la varianza de X_i, si las X_i son independientes y $V(Y)$ es mucho más grande que cualquier componente individual $c_i \cdot V(X_i)$ de una X_i cuya distribución no es normal."

En el caso de una serie de mediciones repetidas (X_1, X_2, … X_n), el Teorema del Límite Central es aplicable a la media $\overline{X}$ donde $c_i = 1/n$

$$\overline{X} = \sum_{i=1}^{n} \frac{1}{n} \cdot X_i \qquad (3)$$

con la particularidad que todas las X_i provienen de la misma distribución con la misma esperanza $E(X_i) = \mu$ y varianza $V(X_i) = \sigma^2$.

Para la esperanza y varianza de $\overline{X}$ resulta:

$$E(\overline{X}) = \sum_{i=1}^{n} \frac{1}{n} \cdot E(X_i) = \mu \qquad (4)$$

$$V(\overline{X}) = \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 \cdot V(X_i) = \frac{\sigma^2}{n} \qquad (5)$$

2. DESARROLLO

Con simulaciones numéricas se muestra para diferentes tipos de distribución de los X_i la aproximación de la distribución de la media $\overline{X}_n$ de n mediciones repetidas a una distribución normal. En hojas de cálculo de Excel se generan 10 000 juegos de datos aleatorios X_{ki}, (i = 1,... 10; k = 1,... 10 000) simulando de esta manera 10 000 ciclos de la medición. Después se calculan las medias para cada uno de los 10 000 ciclos (con índice k)

$$(\overline{X}_n)_k = \frac{1}{n} \cdot \sum_{i=1}^{n} X_{ki} \qquad (6)$$

para n = 2, 3, 5 y 10 y se generan histogramas de frecuencia para ver las distribuciones de los $\overline{X}_n$. Las desviaciones estándar de los $\overline{X}_n$

$$s(\overline{X}_n) = \sqrt{\frac{1}{10\,000} \cdot \sum_{k=1}^{10\,000} \left(\mu - (\overline{X}_n)_k\right)^2} \qquad (7)$$

se comparan con los valores teóricos que se esperan de acuerdo a (5):

$$\sqrt{V(\overline{X}_n)} = \frac{\sigma}{\sqrt{n}} \qquad (8)$$

donde σ es la desviación estándar de los X_i.

3. DISTRIBUCIÓN NORMAL: $X_i \sim N(0, 1)$

En el primer ejemplo se consideran datos X_i provenientes de una distribución normal con $\mu = 0$ y $\sigma = 1$.

La figura 1 muestra la distribución de los datos generados X_i, la figura 2 las distribuciones de $\overline{X}_2$ y $\overline{X}_3$ y la de $\overline{X}_{10}$ y una comparación con una normal. Como se espera, las distribuciones de las $\overline{X}_n$ se parecen a distribuciones normales. Un incremento en n genera distribuciones más estrechas, lo cuál se observa de forma cuantitativa mediante las desviaciones estándar $s(\overline{X}_n)$ calculadas con los 10 000 datos de los $\overline{X}_n$, que siguen la expresión

$$s(\overline{X}_n) \approx \frac{\sigma}{\sqrt{n}} \qquad (9)$$

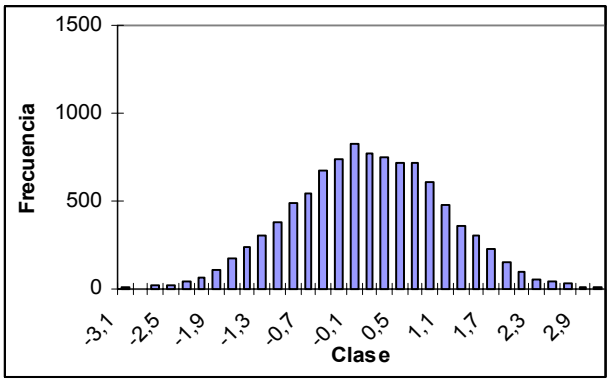
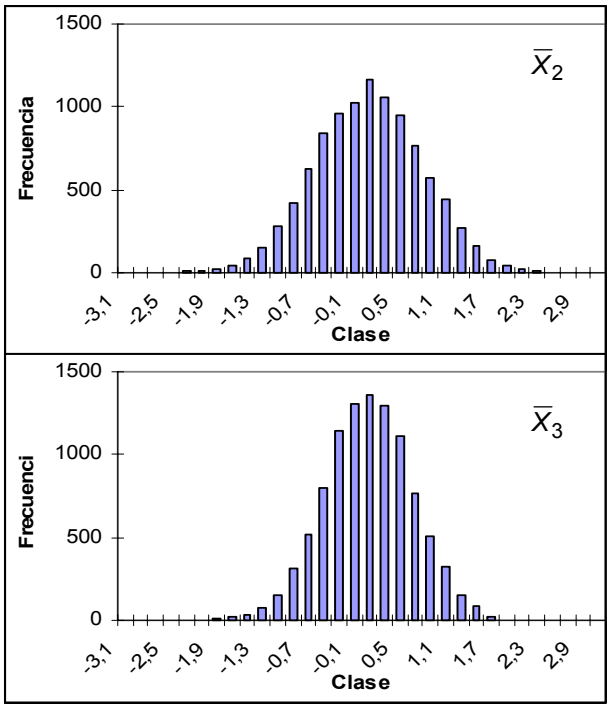


Fig. 1: *Histograma los $X_i \sim N(0, 1)$*



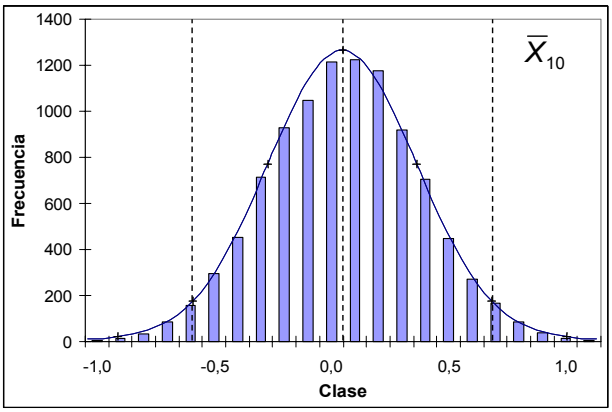


Fig. 2: *Distribución de las medias $\overline{X}_2$, $\overline{X}_3$ y $\overline{X}_{10}$ y comparación con una distribución normal con $\sigma = 1/\sqrt{10} = 0{,}316$ (línea).*

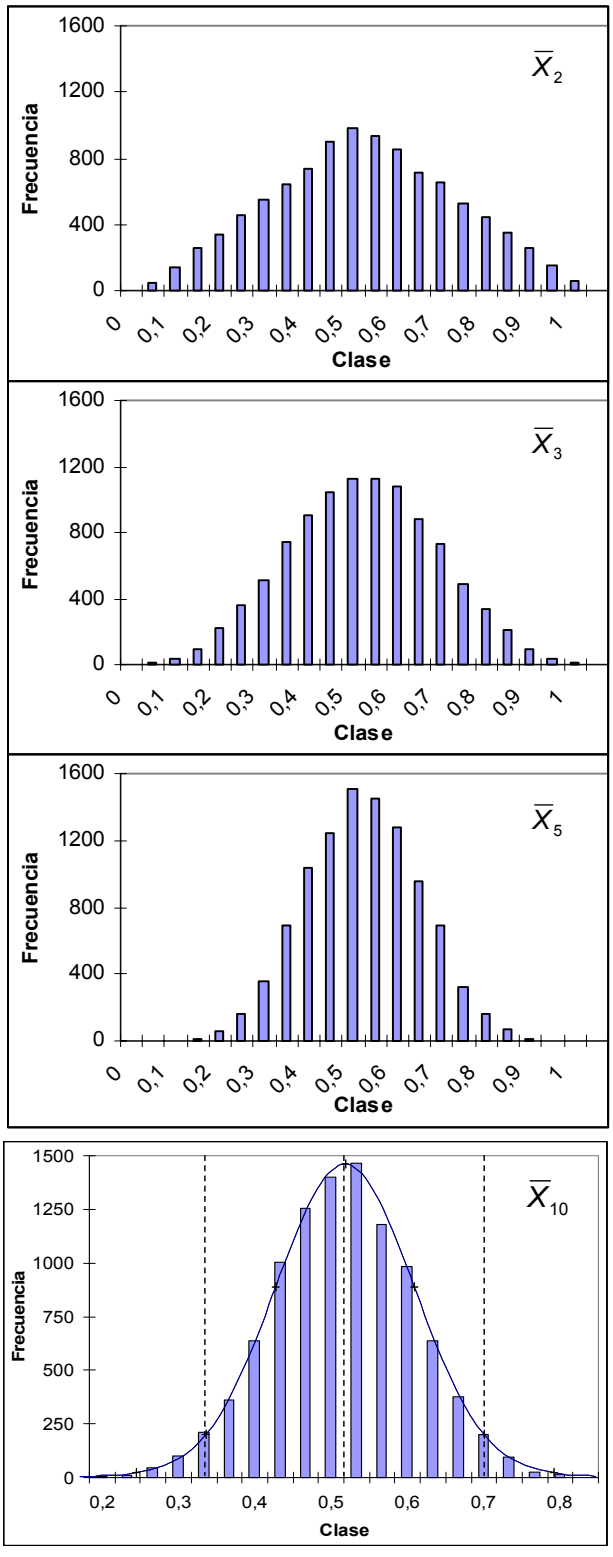


***Fig. 3:**.Distribución de las medias $\overline{X}_2$, $\overline{X}_3$, $\overline{X}_5$ y $\overline{X}_{10}$ en el caso de X_i ~ U(0,1) y la comparación con una normal con $\sigma = 0{,}289/\sqrt{10} = 0{,}091$ (línea).*

4. DISTRIBUCIÓN UNIFORME: X_i ~ U(0, 1)

A continuación se desarrolla el ejercicio con datos de los X_i provenientes de una distribución uniforme entre 0 y 1. La figura 4 muestra la distribución de X_i, la figura 3 las distribuciones de $\overline{X}_2$, $\overline{X}_3$, $\overline{X}_5$ y $\overline{X}_{10}$ y una comparación con una distribución normal (en el caso de $\overline{X}_{10}$). Se observa cómo las distribuciones de las $\overline{X}_n$, incrementando n se acercan cada vez más a distribuciones normales. Igual que en el caso anterior, las desviaciones estándar $s(\overline{X}_n)$ resultan de acuerdo a la expresión (9): $s(\overline{X}_n) \approx 0{,}289/\sqrt{n}$

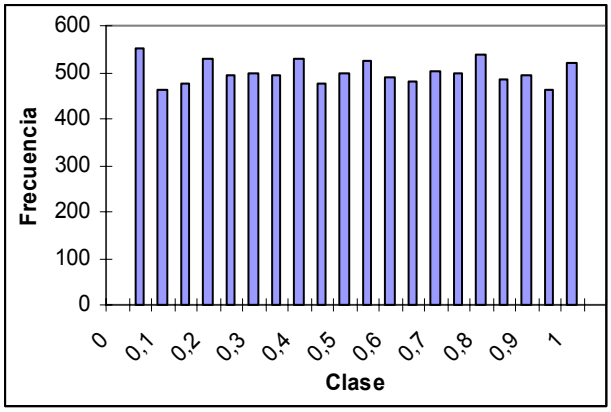


***Fig. 4:** Histograma los X_i provenientes de una distribución uniforme U(0, 1)*

5. DISTRIBUCIÓN BERNOULLI: X_i ~ B(0, 1)

Como siguiente ejemplo se desarrolla el ejercicio con una distribución Bernoulli, que genera los valores 0 y 1 con la misma probabilidad de 50%.

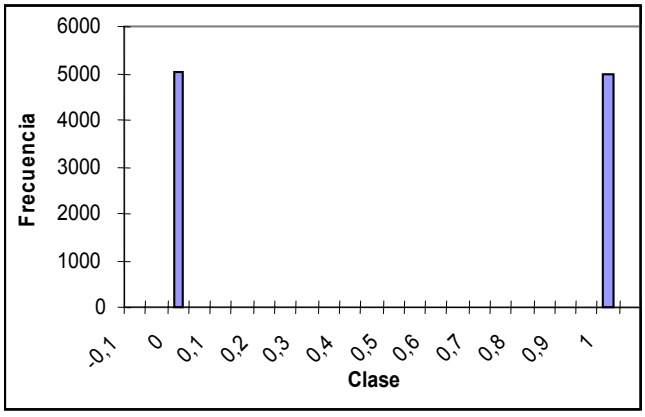


***Figura 5:** Histograma los X_i ~ B(0, 1).*

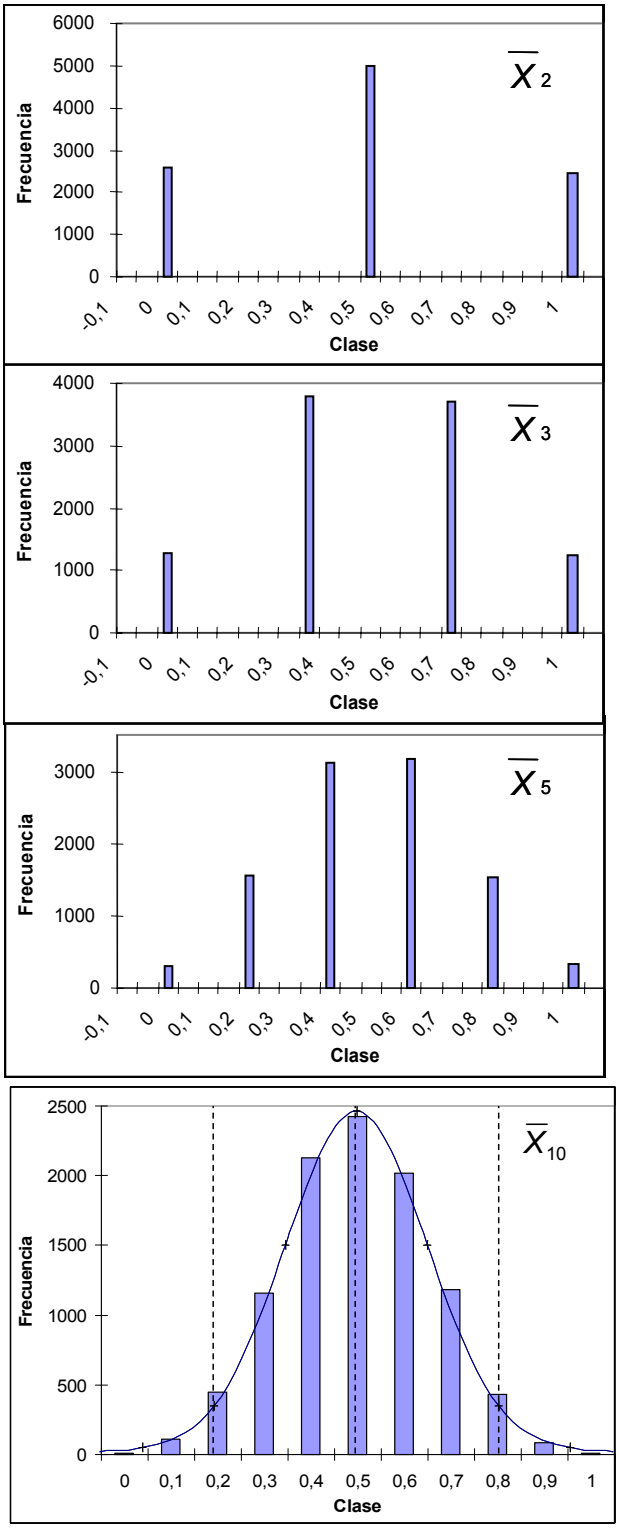


Figura 6: *Distribución de las medias $\overline{X}_2$, $\overline{X}_3$, $\overline{X}_5$ y $\overline{X}_{10}$ en el caso de X_i ~ B(0, 1) y la comparación con una normal con $\sigma = 0{,}5/\sqrt{10} = 0{,}158$ (línea)*

La figura 5 muestra la distribución de X_i y la figura 6 las de $\overline{X}_2$, $\overline{X}_3$, $\overline{X}_5$ y $\overline{X}_{10}$. También en este caso se observa como las distribuciones de las $\overline{X}_n$ se acercan a una normal y que las desviaciones estándar siguen la expresión (9): $s(\overline{X}_n) \approx 0{,}5/\sqrt{n}$.

6. DISTRIBUCIÓN NORMAL CON RESOLUCIÓN FINITA: *X ~ rnd[N(0,25 , 0,5)]*

Como último ejemplo se desarrolla el caso de datos X_i provenientes de una distribución normal y redondeados a enteros $X \sim rnd[N(\mu, \sigma_0)]$, simulando de esta forma la medición de datos con errores aleatorios normalmente distribuidos, utilizando un instrumento con resolución de 1 unidad [1].

Con la selección de los parámetros de $\mu = 0{,}25$ y $\sigma_0 = 0{,}5$ para la distribución normal y el número limitado de 10 000 datos generados resultó en el ejemplo analizado que la generación de los datos se limita a los valores -1, 0, 1, y 2 con una distribución asimétrica a cero. En mediciones reales con un instrumento de medición con una resolución "burda" una situación como está es común.

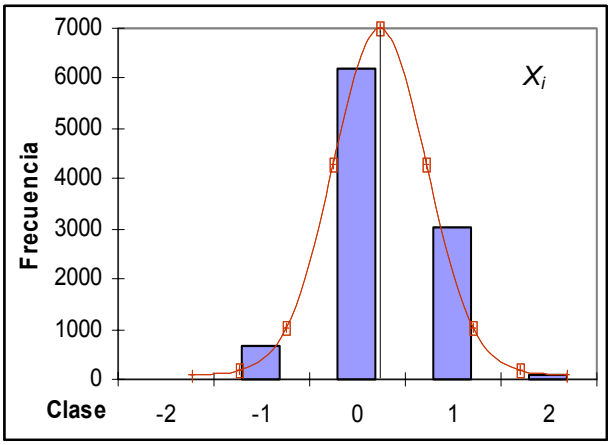


Fig. 7: *Distribución de los X_i provenientes de una distribución normal redondeada $X \sim rnd[N(0{,}25 , 0{,}5)]$ y la función de distribución de la cuál provienen los datos originales (línea).*

La figura 7 muestra la distribución de X_i , la figura 8 las distribuciones de $\overline{X}_2, \overline{X}_3, \overline{X}_5$ y $\overline{X}_{10}$. También en este caso se observa que las distribuciones de los $\overline{X}_n$ se acercan a una distribución normal.

[1] Se usa la notación "***rnd***" para "redondear a enteros".

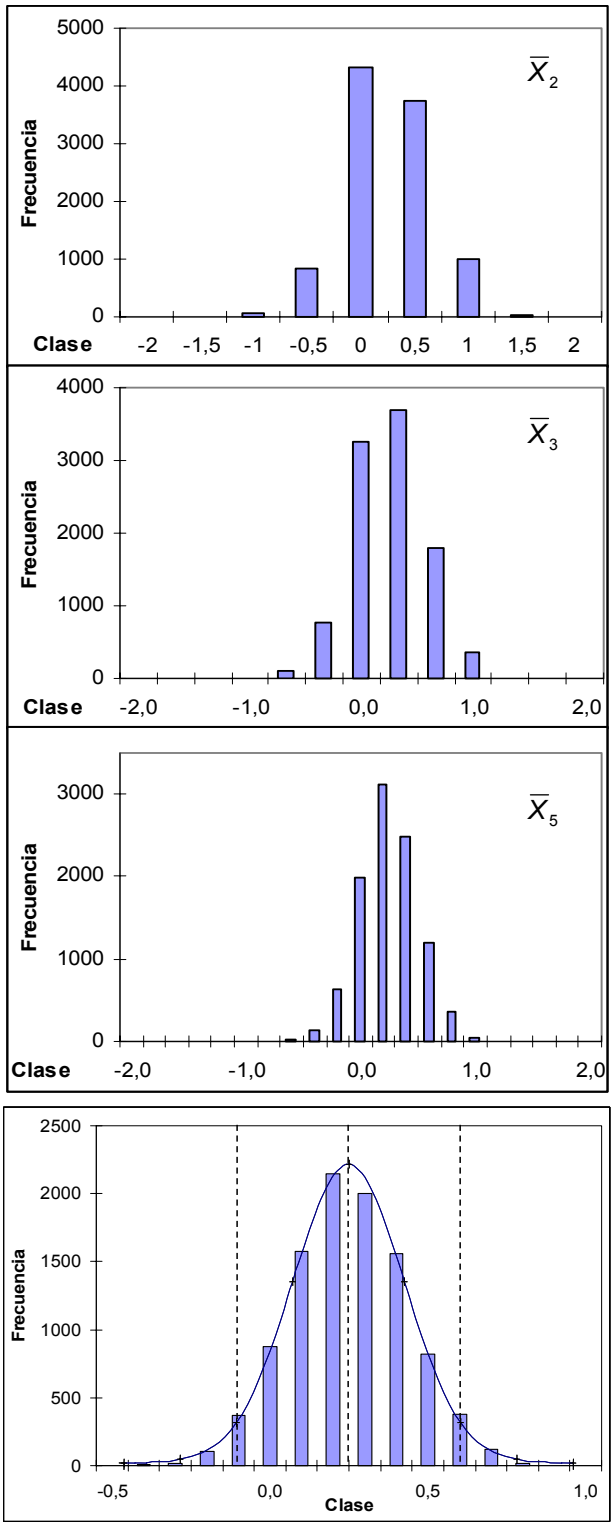


Fig. 8: *Distribución de las medias $\bar{X}_2$, $\bar{X}_3$ y $\bar{X}_5$ (arriba) para X ~ rnd[N(0,25 , 0,5)] y comparación de la distribución de los $\bar{X}_{10}$ (abajo) con una normal con μ = 0,25 y σ = 0,184 (línea)*

En la verificación de la expresión (9) uno se enfrenta con el siguiente problema: tomando como desviación estándar de los X_i el valor $\sigma_0 = 0{,}5$ de la distribución normal, de la cual provienen los datos originales, se obtienen valores más pequeños que las desviaciones estándar de las medias $\bar{X}_n$ obtenidas por la simulación (ver tabla 1). Para cumplir la expresión (9), hay que sustituir σ_0 por la desviación estándar $s(X)$ calculada a partir de los valores X_i generados por el redondeo.

n	2	3	5	10
Media $(\bar{X}_n)$	0,2436	0,2479	0,2491	0,2488
$s(\bar{X}_n)$	0,4093	0,3322	0,2578	0,1841
$\frac{s(X)}{\sqrt{n}}$	0,4088	0,3376	0,2556	0,1831
$\frac{\sigma_0}{\sqrt{n}}$	0,3536	0,2887	0,2236	0,1581

Tabla 1: *Comparación de $s(\bar{X}_n)$ con $s(X)/\sqrt{n}$ y $\sigma_0/\sqrt{n}$ para X ~ rnd[N(0,25 , 0,5)]. $\sigma_0 = 0{,}5$ y s(X) = 0,5781 es la desviación estándar calculada con los datos redondeados obtenidos de la simulación.*

Evidentemente, el redondeo de los datos provenientes de la distribución normal incrementa su desviación estándar. Este efecto puede ser interpretado por la contribución de la discretización o "resolución" a la dispersión de los X_i :

$$X_i = \mu + \delta X_i + \delta R_i \qquad (10)$$

$\delta X \sim$ N(0, σ_0) es la dispersión "original" de los datos y δR la contribución debido al redondeo. Si σ_0 es mayor que la resolución R, la dispersión de los X_i abarca varios intervalos de la resolución y la distribución del error por el redondeo δR se aproxima a una uniforme $\delta R \sim$U(-R/2, R/2) con media $E(\delta R) \approx 0$ y varianza $V(\delta R) = E(\delta R^2) \approx R^2/12$. En consecuencia, la media $E(X)$ es igual a μ y la varianza $s^2(X)$ de los X_i se obtiene mediante:

$$s^2(X) = E\left[(X-\mu)^2\right] = \\ = E\left[(\delta X + \delta R)^2\right] = \\ = E\left(\delta X^2\right) + 2\cdot E(\delta X \cdot \delta R) + E\left(\delta R^2\right) \quad (11)$$

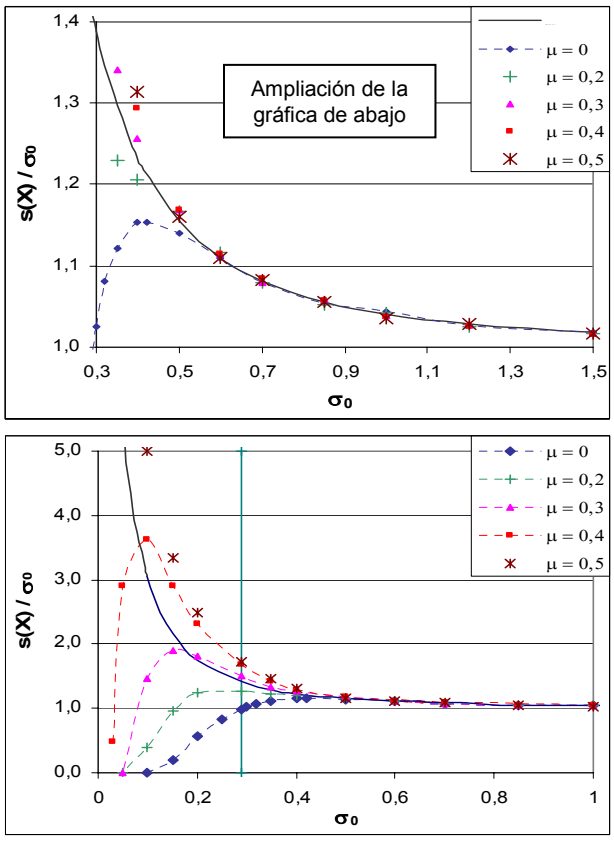


Fig. 9: *Comportamiento de las desviaciones estándar s(X) de los datos redondeados X~rnd[N(μ,σ0)] en relación a σ_0 y la comparación con el comportamiento dado por la ecuación (12).*

Si σ_0 es significativamente mayor que $R/\sqrt{12}$ la distribución de ($\delta X \cdot \delta R$) puede ser considerada como aproximadamente simétrica, resultando en una media $E(\delta X \cdot \delta R) \approx 0$, así que finalmente resulta:

$$s(X) \approx \sqrt{\sigma_0^{\,2} + \frac{R^2}{12}} \quad (12)$$

Este resultado se verificó mediante simulaciones numéricas, generando datos $X \sim rnd[N(\mu,\sigma_0)]$. con diferentes σ_0 y μ . El resultado se muestra en la figura 9. Se observa que para $\sigma_0 > R/\sqrt{12} = 1/\sqrt{12}$ (línea vertical) las desviaciones estándar s(X) siguen en buena aproximación $\sqrt{\sigma_0^{\,2} + R^2/12}$, indicado por la línea, para todos los valores de μ , sin embargo para $\sigma_0 < R/\sqrt{12}$ los valores desvían de este comportamiento, debido a que la aproximaciones hechas en las ecuaciones (11) y (12) no son válidas.

7. CONCLUSIONES

Se ilustró mediante simulaciones con datos aleatorios que, de acuerdo al Teorema del Límite Central, la distribución de la media de un número de *n* datos se aproxima a una distribución normal, independientemente de la distribución de los datos originales y que su desviación estándar disminuye por $\sigma/\sqrt{n}$ de acuerdo a la ecuación (8). Se presentaron estos efectos con datos provenientes de una distribución uniforme, una distribución Bernoulli una distribución normal y una distribución normal con redondeo ("resolución").

Adicionalmente, los resultados muestran que el redondeo de los datos, simulando la resolución de un instrumento de medición, incrementa su desviación estándar de acuerdo a la ecuación (12), por lo menos mientras su desviación estándar sea mayor que la incertidumbre por resolución $R/\sqrt{12}$. Esto indica que la resolución de un instrumento de medición contribuye a la incertidumbre por repetibilidad.

En consecuencia, considerando la incertidumbre por resolución por separado en el presupuesto de incertidumbre, lo cual es una práctica común, lleva a un conteo doble de ella en la incertidumbre combinada. Una práctica más apropiada puede considerar en el presupuesto de incertidumbre solamente la mayor de las dos incertidumbres por repetibilidad o resolución, como lo propone por ejemplo [2].

REFERENCIAS

[1] Guide to the Expression of Uncertainty in Measuremants, BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, OIML, 1995

[2] Geometrical product specifications (GPS) – Inspection by measurement of workpieces and measuring equipment, Part 2, ISO/TR 14253-2