SIMULACIÓN DE LA SUPERPOSICIÓN DE DISTRIBUCIONES DE PROBABILIDAD RECTANGULARES SIMÉTRICAS COMO APLICACIÓN DEL TEOREMA DEL LÍMITE CENTRAL

Nelson Bahamón C – Alexander Martínez L
Instituto Nacional de Metrología (INM)
Bogotá, Colombia
(57) (1) 2 542 222 nbahamon@inm.gov.co amartinez@inm.gov.co

Resumen: La importancia del teorema del límite central en metrología es inmediata, por lo cual la comprensión del mismo es fundamental para quien se está formando en este campo. Un ejemplo sencillo y llamativo que permite aproximarse en forma natural a este concepto, es la superposición de las distribuciones de probabilidad generadas con dados. Se hizo una aplicación en la que se simula dicha superposición y se puede visualizar el resultado de forma clara y pudiendo cambiar fácilmente algunos parámetros. La simulación ha sido útil y exitosa al ser usada por estudiantes y metrólogos del INM.

1. INTRODUCCIÓN.

En este trabajo se aborda el tema del aprendizaje y/o asimilación del concepto del teorema del límite central y su importancia en metrología. [1][2].

El propósito es facilitar el estudio de este concepto mediante un ejemplo que es muy visual e intuitivo; y que facilita la comprensión del mismo por parte de quienes no tienen necesariamente una sólida formación en matemáticas y estadística, como sucede con frecuencia con personas en la industria, que se han ido acercando al tema de la metrología, debido a la difusión que de esta se ha hecho. Por esta razón se han colocado en este artículo algunas explicaciones que pueden resultar muy evidentes para quienes tienen experiencia en el tema.

El ejemplo en cuestión es la superposición de las distribuciones de probabilidad asociadas al lanzamiento de varios dados. Para ello se hizo una simulación basada en el generador de números aleatorios de Excel ®, en realidad pseudoaleatorios.

La GUM (Evaluation of measurement data – Guide to the expression of uncertainty in measurement) [1] es el documento que establece los lineamientos a nivel mundial en todo lo relacionado con metrología. Allí se trata la importancia del teorema del límite central, especialmente en el numeral G2. Es difícil y temerario, establecer un porcentaje pero se podría decir que al menos en el 96% de los casos en que se hace estimación de incertidumbres en una medición, se está aplicando implícita o explícitamente, el teorema del límite central, por considerar que la distribución de salida es aproximadamente normal.

2. SIMULACIÓN.

Para explicar la simulación, se da primero un breve resumen respecto del problema en metrología de estimar la distribución de probabilidad de una variable de salida, dadas las distribuciones de probabilidad de las variables de entrada; luego se hace referencia al teorema del límite central así como su relación con dicho problema; finalmente se explica la realización de la simulación y como se utiliza.

2.1. Distribución de probabilidad de salida.

Sin importar el campo experimental, el problema común en metrología es el siguiente [1]:

Se tiene un mensurando Y que depende de las magnitudes de entrada $X_1, X_2, \dots, X_N$, de forma tal que la relación funcional es conocida, es decir:

$$Y = f(X_1, X_2, \dots, X_N) \quad (1)$$

Se obtiene una estimación y de Y. Mediante los métodos matemáticos pertinentes, que involucran entre otras cosas los coeficientes de sensibilidad c_i asociados a cada variable de entrada X_i, se obtiene la incertidumbre típica combinada u_c de y.

Entonces se quiere estimar un intervalo apropiado para Y que tiene una alta probabilidad o nivel de confianza de que los posibles valores estén contenidos en él.

Este nivel de confianza se caracteriza mediante un porcentaje p mientras que el intervalo, lo caracteriza una nueva incertidumbre U_P, llamada incertidumbre expandida. El intervalo se escribe entonces de la siguiente forma:

$$y - U_P \leq Y \leq y + U_P \quad (2)$$

La incertidumbre expandida se relaciona con la incertidumbre combinada mediante el factor k_p de cobertura, como:

$$U_P = k_p u_C \quad (3)$$

Entonces el problema consiste en hallar k_p para obtener una incertidumbre expandida U_P que genere un intervalo de confianza p. Para ello es muy importante conocer la distribución de probabilidad asociada a Y. [1]

2.2. Teorema del límite central.

La distribución de probabilidad de Y puede obtenerse mediante convolución de las variables de entrada X_i si se cumplen la siguientes dos condiciones [2]:

- Que se conozcan las distribuciones de probabilidad de las variables X_i
- Que Y dependa linealmente de las variables de entrada.

Además de lo complejo que puede resultar realizar dicha convolución, no siempre se puede, porque generalmente no se cumplen plénamente, estas dos condiciones. Entonces se utilizan aproximaciones basadas en el teorema del límite central el cual se enuncia a continuación [3]:

La distribución de Y será aproximadamente normal, y tendrá la esperanza matemática:

$$E(Y) = \sum_{i=1}^{N} c_i E(X_i) \quad (4)$$

y su varianza estará dada por:

$$\sigma^2(Y) = \sum_{i=1}^{N} {c_i}^2 \sigma^2(X_i) \quad (5)$$

donde: $E(X_i) =$ Esperanza matemática de X_i
$\sigma^2(X_i) =$ Varianza de X_i

siempre y cuando se satisfagan las siguientes dos condiciones:

- Que las varialbes de entrada sean independientes entre si.
- Que $\sigma^2(Y)$ sea mucho mayor que cualquier $c_i\sigma^2(X_i)$ de una X_i cuya distribución no sea normal. En otras palabras, que ninguna de las variables de entrada con distribución diferente a la normal, sea dominante.

Si se puede demostrar que se cumplen aproximadamente, las condiciones para la validéz del teorma del límite central, el problema de obtener k_p se simplifica notoriamente. Basta tomar un valor de la distribución normal; por ejemplo, en dicha distribución se tiene que para $p = 95{,}45\%$ $k_p = 2$.

Valdría la pena anotar que el uso del valor $k_p = 2$, es bastante "popular" y utilizado en la práctica de forma sistemática, de manera que se olvida incluso su procedencia. Un mejor método para obtener k_p se hace mediante la obtención del número de grados de libertad [1][2].

2.3. Programa de cómputo.

Se realizó la simulación en el software Excel. Mediante el uso de números aleatorios (pseudoaleatorios), la simulación genera el resultado de arrojar un determinado número de dados un cierto número de veces. Entónces el mensurando es simplemente el valor obtenido en la cara superior del dado después de ser arrojado cuando se trata de un solo dado y es la suma de dichos valores cuando se trata de varios dados.

Si se tiene un solo dado, la distribución de probabilidad es la asociada solo a ese resultado; es decir, se trata de una distribución rectangular, ya que para un solo dado todas las caras tienen la misma probabilidad, (1/6) de ser el valor medido. Cuando se trata de varios dados, la distribución de probabilidad resultante será la superposición de distribuciones de probabilidad rectangulares.

Los parámetros de la simulación son simples. Se puede variar el número de dados (D) entre 1 y 6; y se puede variar el número de lanzamientos (M, "medidas") entre 2 y 10000. Cada vez que se quiere generar un nuevo juego de resultados se oprime F9 en el teclado.

Como resultado, además de una tabla extensa de valores, se muestra una gráfica donde aparece el histograma respectivo y se muestran los ajustes a tres distribuciones; a saber, distribuciones rectangular, triangular y cuadrada. De esta manera,

es posible analizar cual de las distribuciones se ajusta mejor a los datos. La figura 1, muestra el aspecto de la simulación.

Fig. 1. *Aspecto parcial de la simulación. Los parámetros variables son "Medidas" y "# Dados".*

3. RESULTADOS

En la figura 4 (al final), de este documento se muestran los resultados arrojados por la simulación para algunos de los parámetros. Naturalmente los resultados obtenidos son diferentes cada vez que se ejecuta la simulación (**F9**), puesto que se implementan números aleatorios.

Se sugiere ejecutar la simulación[1] con estos u otros parámetros para ver progresivamente los resultados y el funcionamiento de la misma.

Los parámetros utilizados en los resultados mostrados fueron M = 100 y 10000; y para D = 1, 2, 4 y 6.

4. DISCUSIÓN

El propósito de la simulación es descubrir de manera natural lo que se observa cuando aumenta el valor de cada uno de los parámetros (número de dados y número de medidas). Vale la pena aclarar que algunas de las presentes observaciones se hacen analizando los resultados de la simulación, con muchos mas parámetros que los presentados en la figura 3.

Es interesante observar el resultado para un solo dado. Se puede ver claramente como la distribución converge a una forma rectangular entre mayor sea el número de medidas, siendo prácticamente perfecta cuando M = 10000.

En el campo de la probabilidad, esto se conoce como la "ley de los números grandes"; en palabras sencillas lo que se tiene es que la distribución de probabilidad esperada teóricamente así cómo su media, se reproducen mejor en la práctica, entre mas grande sea el número de la muestra. En principio se reproducirían de forma exacta si se pudiera tener un número infinito de datos.

En este punto se le insiste al lector, en que "juegue" con la simulación para comprobar dichos comportamientos.

Por ejemplo, es claro que para un solo dado la distribución de probabilidad teórica es rectangular, porque la probabilidad de que se obtenga el resultado de cada cara es igual y corresponde a 1/6. Al probar con un solo dado y 5 medidas, puede observarse para muchos ensayos (oprimiendo F9), lo diferente del resultado en cada caso, siendo muy extraño que en algún momento se obtenga la esperada distribución rectangular. Un análisis similar se puede hacer respecto del valor esperado (promedio); claramente este es de 3.5, pero hay una gran variabilidad al ejecutar la simulación. Se puede observar como cambia esto, al aumentar gradualmente el número de medidas. Cuando se tiene un dado y 10000 medidas, el comportamiento rectangular de la distribución es muy estable y el promedio rara vez se aleja mas de dos centésimas, de 3.5.

Análisis similares se pueden hacer al cambiar el número de dados. Con dos dados, y 10 medidas, no es posible detectar un comportamiento definido; con 100 medidas tampoco, pero si se percibe que la distribución no es rectangular; hay un cierto decaimiento de las colas, y un aumento de la probabilidad hacia el centro; con 1000 medidas empieza a notarse que el comportamiento de la distribución es triangular y con 10000 se hace mas evidente.

Si bien el estudiante puede comprobar este comportamiento triangular de la distribución, mediante la ejecución de la simulación, valdría la pena que se pregunte: "¿Es esta la distribución que se esperaría teóricamente?" La respuesta es si. La comprobación es fácil y también muy didáctica; basta que se estudie el número de microestados asociado a cada resultado. Por ejemplo, para la ocurrencia del valor 5, se tienen 4 posibles microestados, como se ilustra en la figura 2.

[1] Si el lector requiere el archivo para efectuar la simulación la puede solicitar a los autores por correo electrónico.

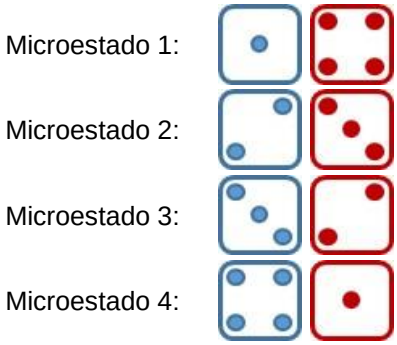


Fig. 2. Microe*stados asociados al valor 5*

Al comprobar todos los microestados posibles asociados a cada resultado para dos dados (es decir de 2 a 12), se encuentra claramente que la distribución teórica esperada de probabilidad, es de tipo triangular. Esta demostración es perfectamente válida y se extrae del campo de la mecánica estadística. Al hacer el cálculo completo se obtiene la siguiente gráfica con su respectiva tabla de datos:

Macroestado	Microestados
2	1
3	2
4	3
5	4
6	5
7	6
8	5
9	4
10	3
11	2
12	1

Fig. 3. *Caso de solo dos dados. La distribución de probabilidad teórica esperada es triangular.*

Para el caso de tres dados puede notarse que la distribución se aproxima a una forma gaussiana cuando se aumenta el número de medidas; dicho comportamiento parece ser mas evidente, en la medida en que el número de dados aumenta.

Entonces se ve con mucha claridad que se está cumpliendo el teorema del límite central; entre mayor sea el número de medidas y mayor sea el número de dados, la convolución de las distribuciones rectangulares dadas por cada dado, da como resultado una distribución normal.

De todas formas hay mas aspectos que se pueden revisar, si se es detallista con la simulación. Por ejemplo, para 10000 medidas, puede verse que con tres y cuatro dados, la parte central del histograma siempre se encuentra por debajo de la curva normal. Esto sugiere que la curva puede ser realmente una t de Student, la cual tiende a la normal cuando el número de grados de libertad es muy grande. El mismo comportamiento se observa para cinco y seis dados, aunque es menos notorio. Lo anterior está en plena concordancia con la información teórica estadística al respecto. La distribución t de Student, describe fenómenos aleatorios con una tendencia central similar a la gaussiana, pero la primera se ajusta mejor cuando se tiene una muestra pequeña. Cuando la muestra tiende hacia infinito, la distribución t, tiende a la gaussiana. Se puede hacer esta discusión mas exhaustiva, pero se sale del alcance de este trabajo.

Se ha hecho la discusión teniendo fijo el número de dados y aumentando el de medidas. Se puede hacer también fijando el número de medidas y variando el de dados.

5. CONCLUSIONES

Se hizo una simulación que ilustra didácticamente el teorema del límite central. Como se comprobó con estudiantes, la simulación es de gran utilidad para comprender este concepto y su papel dentro del proceso de estimación de incertidubres en metrología.

REFERENCIAS

[1] J.C.G.M. Evaluation of Measurement Data – Guide to the Expression of Uncertainty in Measurement. Working Group 1 of the Joint Committe for Guides in Metrology. First edition – September 2008.

[2] A. E. Fridman. The Quality of Measurements a Metrological Reference. Politechnic Institute Russia. Springer 2012.

[3] S. V. Gupta. Measurement Uncertaninties Physical Parameters and Calibration of Instruments. Springer-Verlag Berlin Heidelberg 2012.

[4] C. F. Dietrich. Uncertainty Calibration and Probability the Statics of Scientific and Industrial Measurement. Taylor & Francis Group. Second Edition 1991.

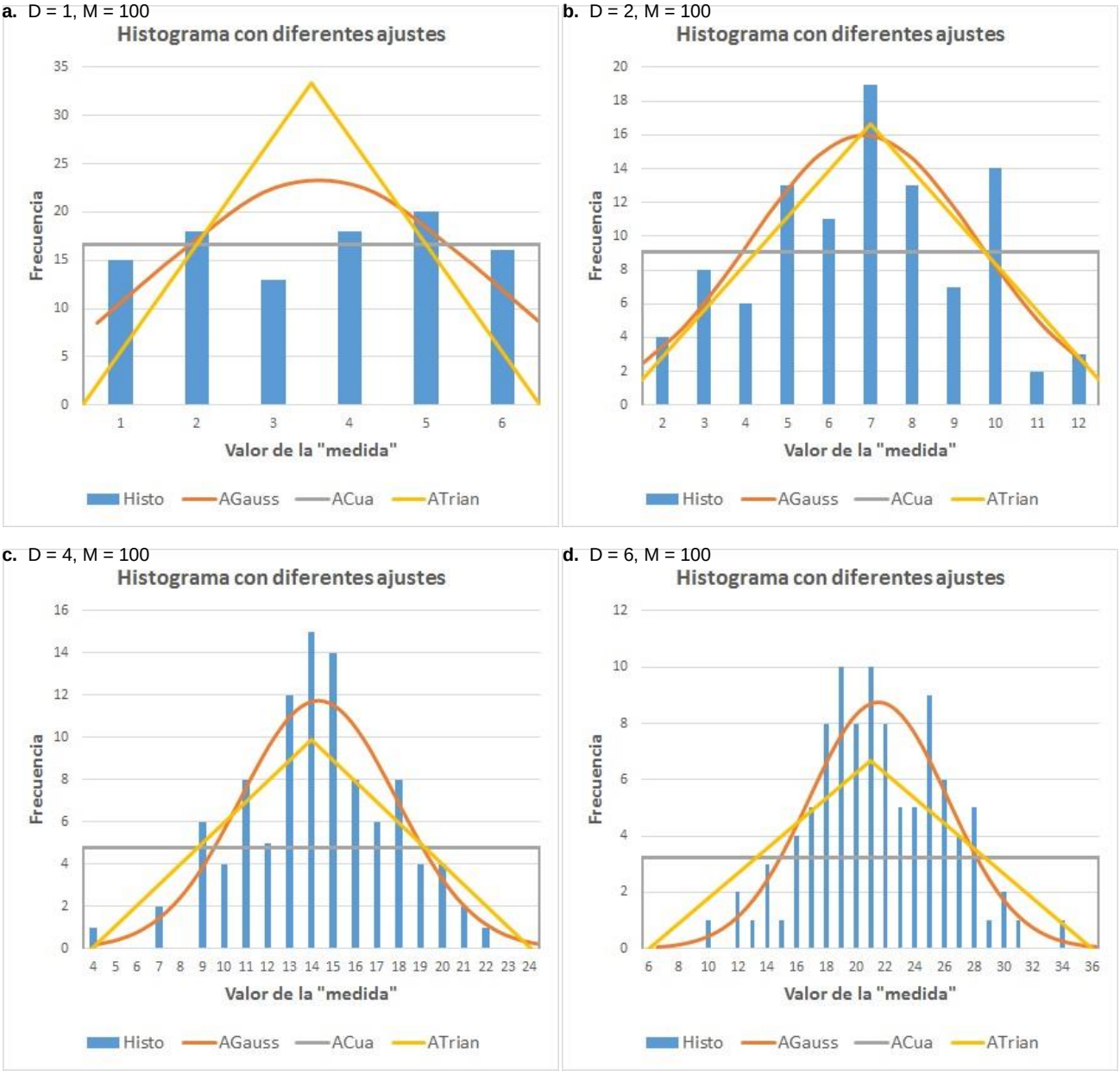
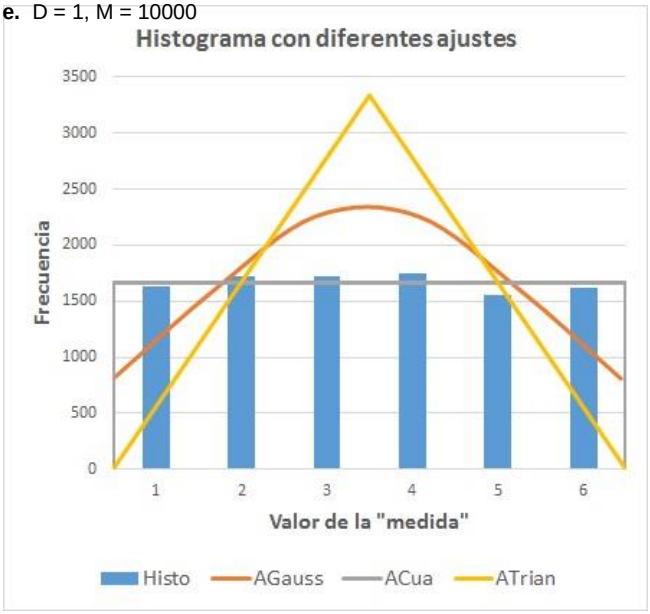
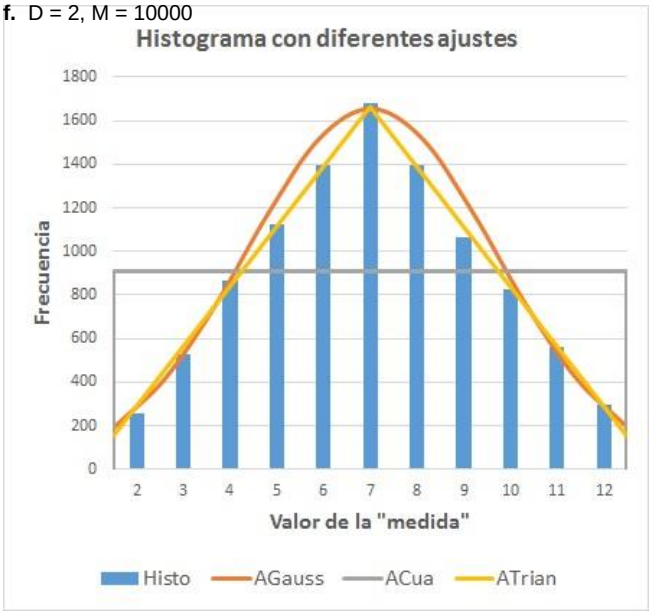


Fig. 4a – 4d. *Resultados de la simulación con diferentes parámetros de # Dados (*D*)" y "Medidas (*M*)"*

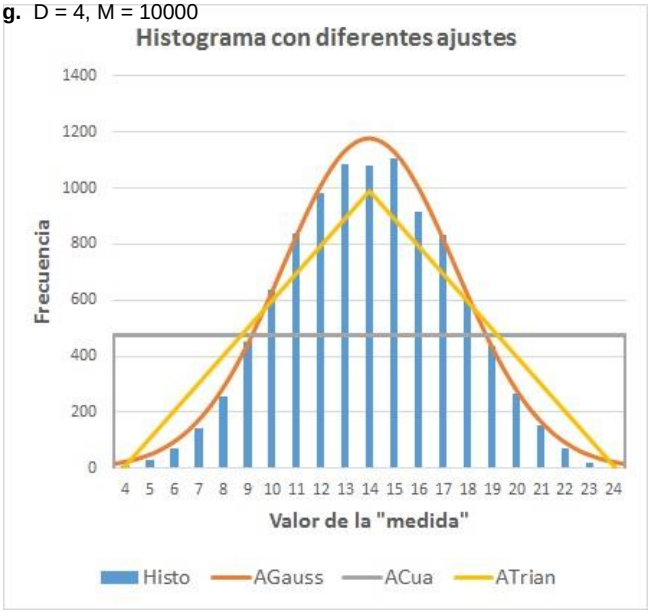
e. D = 1, M = 10000



f. D = 2, M = 10000



g. D = 4, M = 10000



h. D = 6, M = 10000

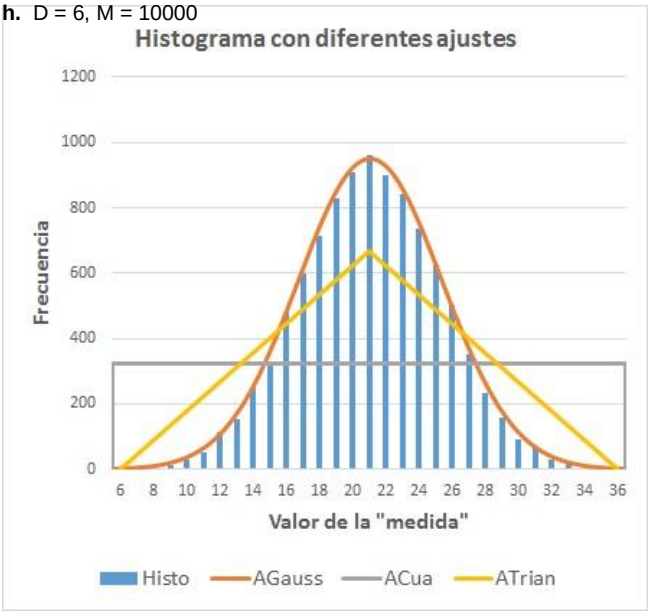


Fig. 4d-4h. *Resultados de la simulación con diferentes parámetros de # Dados (*D*)" y "Medidas (*M*)"*