Revista Colombiana de Estadística
Vol. **22** No.2 (1999) *págs 27 - 38*

Una aproximación a la distribución de la varianza en poblaciones simétricas no normales

JESÚS ANTONIO AVILA G.
JAIRO ALFONSO CLAVIJO M. *

Resumen.

En este artículo se hacen algunas consideraciones acerca de la distribución del estimador de la varianza para algunas distribuciones simétricas continuas no normales, proponiendo para ello una función de densidad que permita estimar sus cuatro primeros momentos. Se comparan los resultados con la función de aproximación de Box y se utiliza el método de un sistema de ecuaciones lineales.

Palabras Claves: Distribuciones simétricas, aproximación de Box.

1. Introducción

Considérese una población definida por una variable aleatoria X con distribución normal, de media μ y varianza σ^2. Si $\{X_1, X_2, \ldots, X_n\}$ es una muestra aleatoria de X (es decir, X_i está idénticamente distribuida con X para todo i y X_i, X_j son independientes) se sabe que $\bar{X} = \frac{1}{n}\sum X_i$ y $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$ son respectivamente estimadores insesgados de μ y σ^2. En este caso se sabe también que $\bar{X}$ tiene distribución normal con media μ y varianza $\frac{\sigma^2}{n}$. Además la variable aleatoria $Y = \frac{(n-1)S^2}{\sigma^2}$ tiene distribución ji-cuadrado con $n-1$ grados de libertad.

Existen otros dos casos en que se conoce exactamente la distribución de la variable aleatoria KS^2 (siendo K una constante distinta de cero), a saber:

*Profesor, Departamento de Matemáticas, Universidad del Tolima, Colombia

1- Cuando la población base es una mezcla de dos normales con distinta media, con función de densidad:

$$f(x) = p\phi_1(x;\mu_1,\sigma^2) + (1-p)\phi_2(x;\mu_2,\sigma^2)$$

con $0 \leq p \leq 1$, siendo $\phi_i(x;\mu_i,\sigma^2)$ la función de densidad de la distribución normal con media μ_i y varianza σ^2. En tal caso la distribución de la variable aleatoria.

$$Y = \frac{(n-1)S^2}{\sigma^2}$$

es

$$h(y) = \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} J(y; n-1, \eta_j^2)$$

donde $J(y; n-1, \eta_j^2)$ es una distribución ji-cuadrado no central con $n-1$ grados de libertad y parámetro de no centralidad dado por $\eta_j^2 = j(n-j)(\mu_1-\mu_2)^2/n\sigma^2$ Tan, Wong(1977).

2- Cuando la población base es una mezcla de dos normales de igual media y distinta varianza, con función de densidad:

$$f(x) = p\phi(x;\mu,\sigma_1^2) + (1-p)\phi(x;\mu,\sigma_2^2)$$

con $0 \leq p \leq 1$, $\sigma_1^2 > 0$, $\sigma_2^2 > 0$, $-\infty < \mu < \infty$ y $\phi(x;\mu,\sigma_2^2)$ es la distribución normal de media μ y varianza σ_i^2, entonces, para $\sigma_1^2 = 1$ y $\sigma_2^2 = \sigma^2$ se tiene que la distribución acumulada de la variable aleatoria $Y = (n-1)S^2$, está dada por

$$Pr(Y \leq t) = \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} \times Pr(\sum \lambda_i Q_j \leq t)$$

donde $\sum \lambda_j Q_j$ es una forma cuadrática en $i+1$ variables normales distribuidas independientemente Mudholkar, Trivedi (1981).

Si la distribución de la población no es normal, se hace uso del teorema del límite central, aumentado el tamaño de la muestra –lo que en algunos casos no es fácil o no es económico– o se hace una transformación de los datos para lograr un mejor ajuste a la distribución normal (esto se utiliza en algunas técnicas de laboratorio en la industria, para determinar la varianza de repetibilidad y de reproducibilidad). Sin embargo no en todos los casos se puede lograr el ajuste deseado.

Lo anterior sugiere estudiar una aproximación a la distribución de S^2 para poblaciones no normales, ya que ella es evaluable exactamente sólo en los tres casos ya citados. Entre todo el conjunto de poblaciones no normales, se encuentran las distribuciones simétricas continuas, que por sus propiedades de convergencia, momentos, etc., facilitan el estudio que se propone. Por tanto, restringiremos el presente artículo a tales poblaciones.

2. Antecedentes teóricos

Ya que no se puede evaluar la distribución exacta de S^2 cuando las poblaciones son *distintas a la distribución normal* o *mezcla de distribuciones normales*, se han generado distintas aproximaciones, basadas fundamentalmente en la aproximación de los momentos teóricos de la distribución de KS^2. Lo anterior se hace puesto que si dos funciones de densidad determinan la misma función generatriz de momentos, entonces, ellas coinciden (Mood, Graybill, Boes (1974)). Por tanto, se busca es una aproximación al mayor número de momentos posibles. Siguiendo esta metodología, Box (1953) aproximó los dos primeros momentos (media y varianza), utilizando una distribución gama. El mostró que si X es una variable aleatoria con función de distribución $f(x)$ y si $Y = \frac{(n-1)S^2}{C_2}$, con $C_2 = Var(X)$, entonces la distribución acumulada de Y se aproxima por:

$$Pr(Y \leq t) \approx \frac{1}{\Gamma(b)\rho^b} \int_0^t y^{b-1} e^{-y/\rho} dy,$$

donde $\rho = Var(Y)/m$, $b = m/r$ y $m = E(Y) = n - 1$ (Mudholkar y Trivedi (1981)).

Lo anterior lleva a la construcción (Método de Box) de una función de densidad de la forma:

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}$$

El mismo Box, estimó los parámetros α y β como:

$$\hat{\alpha} = \frac{n}{\gamma - \frac{n-3}{n-1}} \qquad \hat{\beta} = \frac{\sigma^2}{n}\left(\gamma - \frac{n-3}{n-1}\right)$$

valores que se obtienen como solución de las ecuaciones:

$$\alpha\,\beta = \sigma^2 \qquad \alpha\beta^2 = \frac{\sigma^4}{n}\left(\gamma - \frac{n-3}{n-1}\right)$$

Donde los primeros miembros de las anteriores igualdades representan la media y la varianza de la distribución gama y los segundos la media y la varianza teóricas de la distribución de S^2.

Barton (1953) obtuvo una aproximación por series de Laguerre para la distribución de una suma de cuadrados. De ésta, Roy y Tiku (1962) derivaron una aproximación similar para la distribución de $Y = \frac{(n-1)S^2}{2C_2}$, la cual se enuncia así Mudholkar y Trivedi (1981):

Si X es una variable aleatoria con función de densidad $f(x)$ y $Y = \frac{(n-1)S^2}{2C_2}$, donde $C_2 = Var(X)$, entonces la distribución acumulada de Y está dada por:

$$Pr(Y \leq t) = \int_0^t P_m(y) \sum_{j=0}^{k} a_j^{(m)} L_j^{(m)}(y) dy$$

donde

$$P_m(y) = \frac{1}{\Gamma(m)} y^{m-1} e^{-y}, y \geq 0, \qquad a_j^{(m)} = \Gamma(m) \sum_{i=0}^{j} \binom{j}{i} E(-y)^i / \Gamma(m+i).$$

$$L_j^{(m)}(y) = \frac{1}{j!} \sum_{i=0}^{j} \binom{j}{i} (-y)^i \Gamma(m+j)/\Gamma(m+i),$$

$m = E(Y)$, k es el número de términos en la aproximación y los a_j son constantes determinadas por los primeros j momentos de Y.

Como puede observarse en la aproximación anterior, la función de distribución de KS^2 está dada por una exponencial multiplicada por una suma de polinomios de Laguerre, cuyas constantes dependen de los momentos teóricos de S^2. Aunque se puedan aproximar los j momentos, el proceso para encontrar explícitamente la función de densidad no es práctico debido a los largos y tediosos cálculos numéricos que exigen recursos computacionales.

Tang y Wong (1977) retomaron el trabajo realizado por Box (1953), Roy y Tiku (1962), mezclando ambas teorías, propusieron siguiente aproximación alternativa:

Si $Y \geq 0$ es una variable aleatoria, con $E(Y) = m$ y $E|Y^k| < \infty$ para $k > 2$, entonces la k-ésima aproximación $\Psi_k(y)$ para la distribución $h(y)$ de Y está dada por:

$$\Psi_k(y) = Q_b(y) \sum_{r=0}^{k} c_r L_r^{(b)}(y/\rho)$$

donde

$$\rho = \frac{Var(Y)}{m}, \quad b = \frac{m}{\rho}, \quad Q_b(y) = \frac{1}{\Gamma(b)\rho^b} y^{b-1} e^{-y/\rho}, \text{ y}$$

$$d_r = \binom{b+r-1}{r} c_r = E(L_r^{(b)}(y/\rho))$$

$$= \frac{1}{r!} \sum_{j=0}^{r} \binom{r}{j} (-1)^j E\left(\frac{Y}{\rho}\right) (\Gamma(b+r)/\Gamma(b+j)),$$

para $r = 0, 1, 2, \ldots$, $(d_0 = c_0 = 1,\ c_1 = c_2 = 0)$.

De lo anterior se deduce que la función de distribución de $Y = \frac{(n-1)S^2}{2C_2}$ es una gama multiplicada por una suma de polinomios de Laguerre, cuyas constantes dependen de los momentos teóricos de S^2. Observando los primeros términos de esta aproximación, puede encontrarse que es una generalización de la aproximación de Box, que para $k \geq 4$, produce mejores resultados que los obtenidos por Box.

3. Aproximación Propuesta

Dado que si dos funciones de densidad tienen los mismos momentos, entonces, dichas funciones son iguales. Por lo anterior se han generado las distintas aproximaciones estudiadas, buscando que sus momentos se aproximen a los de la distribución de KS^2. Como se ha visto, las aproximaciones citadas contienen una distribución gama, que depende de dos parámetros. Basados en esto, proponemos una función de densidad -que es el "promedio" de dos densidades gama-, y que depende de cuatro parámetros. Con lo anterior se logra una función cuya forma se asemeja a la distribución deseada y cuyos cuatro primeros momentos se obtienen como solución de un sistema de ecuaciones lineales. La función propuesta es:

$$f(x) = \frac{x^{\alpha_1-1}e^{-\frac{x}{\beta_1}}}{2\beta_1^{\alpha_1}\Gamma(\alpha_1)} + \frac{x^{\alpha_2-1}e^{-\frac{x}{\beta_2}}}{2\beta_2^{\alpha_2}\Gamma(\alpha_2)}$$

donde α_1, α_2, β_1 y β_2 son las soluciones del siguiente sistema de ecuaciones:

$\frac{1}{2}(\alpha_1\beta_1 + \alpha_2\beta_2) = \sigma^2$

$\frac{1}{4}\left[(\alpha_1\beta_1 + \alpha_2\beta_2)^2 + 2(\alpha_1\beta_1^2 + \alpha_2\beta_2^2)\right] = \frac{\sigma^4}{n}\left(\gamma - \frac{n-3}{n-1}\right)$

$\frac{1}{2}\left[(\alpha_1(\alpha_1+1)(\alpha_1+2)\beta_1^3 + \alpha_2(\alpha_2+1)(\alpha_2+2)\beta_2^3\right] - \frac{3}{4}(\alpha_1\beta_1 + \alpha_2\beta_2)$

$\left[\alpha_1(\alpha_1+1)\beta_1^2 + \alpha_2(\alpha_2+1)\beta_2^2\right] + \frac{1}{4}(\alpha_1\beta_1^2 + \alpha_2\beta_2^2)^3 = \frac{\mu_6}{n^2} - \frac{n\mu_2^3(3\gamma-2)}{(n-1)^3} + o\left(\frac{1}{n^3}\right)$

$\frac{1}{2}\left[\alpha_1(\alpha_1+1)(\alpha_1+2)(\alpha_1+3)\beta_1^4 + \alpha_2(\alpha_2+1)(\alpha_2+2)(\alpha_2+3)\beta_2^4\right] -$

$(\alpha_1\beta_1 + \alpha_2\beta_2)\left[\alpha_1(\alpha_1+1)(\alpha_1+2)\beta_1^3 + \alpha_2(\alpha_2+1)(\alpha_2+2)\beta_2^3\right] + \frac{3}{4}$

$(\alpha_1\beta_1 + \alpha_2\beta_2)^2[\alpha_1(\alpha_1+1)\beta_1^2 + \alpha_2(\alpha_2+1)\beta_2^2] - \frac{3}{16}(\alpha_1\beta_1 + \alpha_2\beta_2)^4 = \frac{3n^2\mu_2^4(\gamma-1)^2}{(n-1)^4}$

$+o\left(\frac{1}{n^3}\right)$

debido a su complejidad, el anterior sistema de ecuaciones no puede ser solucionado de la manera usual, ya que difícilmente cabe esperar soluciones racionales, siendo necesario encontrar soluciones reales aproximadas. Es por ello que se aplica el método de mínimos cuadrados, encontrando la solución que hace que la suma de cuadrados de los errores sea mínima.

A través del "Solver" de Excel, se determinaron los coeficientes $\alpha_1, \alpha_2, \beta_1$ y β_2. Este procedimiento modifica los valores de los parámetros a estimar, hasta que la suma de cuadrados de los errores sea lo más cercana a cero posible, teniendo en cuenta los términos principales de los momentos tercero y cuarto de S^2. Cabe anotar que el tercer momento no concuerda con lo propuesto por Cramer (1950), ni con los resultados de simulación, por lo cual se tomaron muestras de tamaño 10000 para estimarlo. De esta manera las soluciones obtenidas son aproximadas.

4. Metodología de simulación

Se seleccionaron las siguientes poblaciones base: beta simétricas de parámetros 2, 4 y 6, parabólica cóncava y convexa, uniforme y triangular superior e inferior.

Para los procesos de simulación, se tomaron intervalos entre -4 y 4. No se pudo muestrear la beta de parámetro superior al 6, pues el factor inicial en la función de densidad acumulada alcanza el orden de 10^4, trayendo consigo problemas en la obtención de las muestras.

Los programas utilizados para las simulaciones fueron modificaciones de los propuestos por Clavijo (1995), donde el procedimiento de simulación -descrito allí mismo- consta fundamentalmente de las siguientes partes:

1- f es la función de densidad definida en el intervalo $[-a, a]$, de una variable aleatoria X, con función de distribución acumulada F.

2- Se divide el intervalo soporte en n subintervalos de longitud $\frac{2a}{n}$, obteniéndose la partición $\{x_0 = -a, x_1, x_2, \ldots, x_n = a\}$.

3- Se calculan los valores $F(x_0) = 0,\ F(x_1), \ldots, F(x_n) = 1$.

4- Si se desea tomar una muestra de tamaño m, $\{y_1, y_2, \ldots, y_m\}$, entonces, se generan m números aleatorios $p_1, p_2, \ldots, p_m$ entre 0 y 1. Los que pueden ser tomados como valores de probabilidad.

5- Para un valor p_k, del paso anterior, se hacen comparaciones con los valores conocidos $F(x_0) = 0,\ F(x_1), \ldots, F(x_n) = 1$, hasta obtener un índice l, tal que, $F(x_l) < p_k \leq F(x_{l+1})$, siendo x_{l+1} el valor y_k de la muestra.

5. Resultados y conclusiones

5.1 Las figuras 1 a 6 muestran la aproximación dada por Box, la aproximación propuesta por nosotros y el histograma de la muestra simulada para diferentes distribuciones simétricas.

5.2 Puede observarse que para valores pequeños de n, la aproximación de Box presenta un pico respecto a la aproximación propuesta, sin embargo esta última, se ajusta más a la forma del histograma. Podría decirse que al aumentar el tamaño de muestra las dos gráficas casi coinciden (tablas 1 a 3). En las gráficas aparece el caso $n = 5$.

5.3 En las tablas 1 a 3 puede observarse que la función propuesta por nosotros genera aproximadamente los cuatro primeros momentos de la distribución de S^2: media, varianza, tercero y cuarto momentos, implicando que al ser utilizada, proporciona mejores resultados en las estimaciones por intervalo. También puede observarse como al aumen-

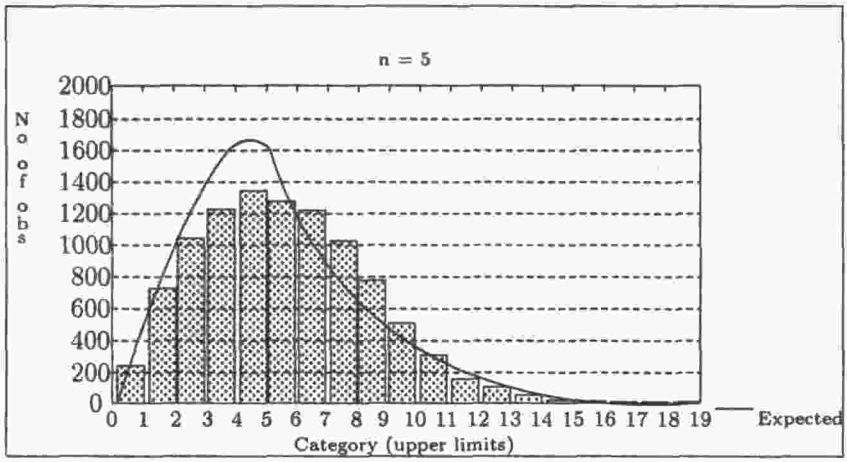
tar el tamaño de muestra, los momentos generados por la aproximación de Box van acercándose a los valores teóricos.

5.4 Se ha obtenido una función de densidad para la distribución de S^2 en poblaciones simétricas continuas, que a diferencia de la obtenida por Box, aproxima los primeros cuatro momentos y aunque no es tan general como las propuestas por Tan-Wong y Roy-Tiku, la facilidad de conceptualización y obtención de los parámetros, hacen de ésta, una herramienta de rápida aplicación en la práctica, sin descartar que el método podría generar funciones que aproximan un mayor número de momentos. Además el proceso seguido para la obtención de la función de densidad propuesta, no tuvo en cuenta ninguna condición sobre las poblaciones base, por lo tanto esta función y el procedimiento seguido podría ser aplicado cuando la población base sea distinta de la normal, sin importar si es simétrica o no.

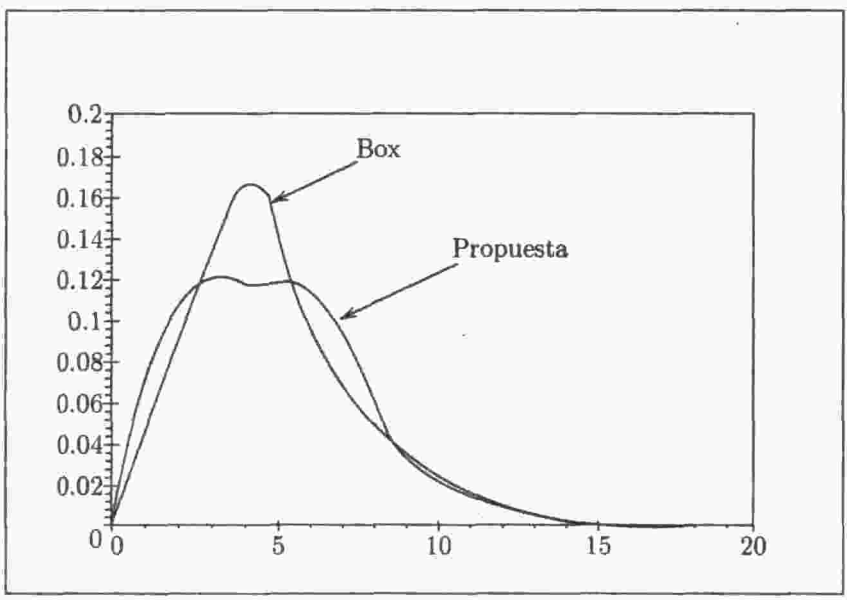
Referencias

1. Mudholkar G.S. y Trivedi M.C., *A Gaussian Approximation to the Distribution of the Sample Variance for Nonormal Populations*, JASA **76** no. 374 (1981).
2. Tan W. Y. y Wong S. P., *On the Roy Tiku Approximation to the Distribution of the Sample Variances from Nonormal Universes*, JASA **72** no. 360 (1977).
3. Clavijo J. A., *Tamaños de Muestra para Distribuciones Simétricas Acotadas*, Tesis de Maestría Universidad Nacional de Colombia (1995).
4. Mood A. M., Graybill F. A., BOES D. C., *Introduction to the theory of statistics*, Mc-Graw-Hill. Tercera Edición, Singapore, 1974.
5. Freund J. E., Walpole R. E.;, *Estadística Matemática con Aplicaciones*, Prentice-Hall Hispanoamericana S. A., México, 1990.
6. Cramer H., *Métodos Matemáticos de la Estadística*, Princeton University Press, 1950.

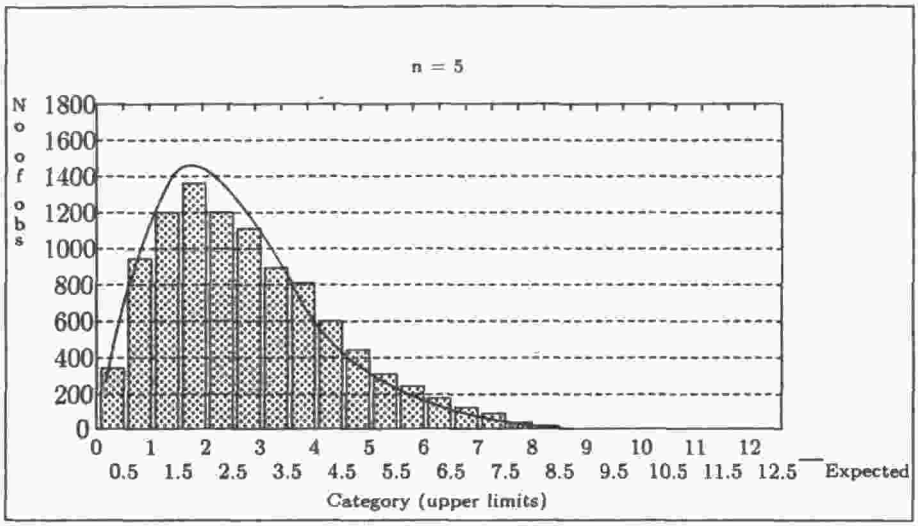
Gráfica 1. Histograma de la muestra simulada y aproximación de Box para una distribución uniforme en [-4, 4] (tamaño $n = 5$).



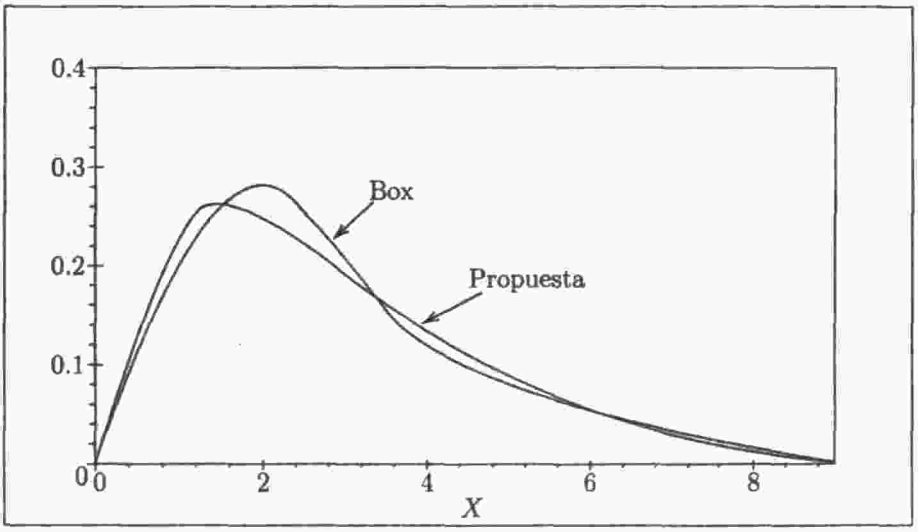
Gráfica 2. Comparación de la aproximación propuesta con la de Box, en una distribución uniforme en $[-4, 4]$ (tamaño de muestra $n = 5$).



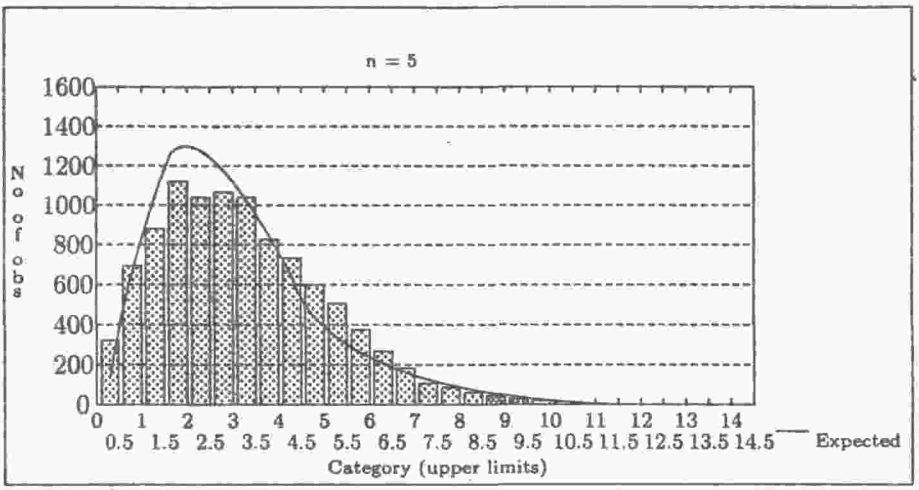
Gráfica 3. Histograma de la muestra simulada y aproximación de Box para una distribución triangular superior pura en [-4, 4] (tamaño $n = 5$).



Gráfica 4. Comparación de la aprox. propuesta con la aproximación de Box, en una distribución triangular superior pura en $[-4, 4]$ (tamaño de muestra $n = 5$).



Gráfica 5. Histograma de la muestra simulada y aproximación de Box para una distribución parabólica convexa en [-4, 4] (tamaño $n = 5$).



Gráfica 6. Comparación de la aproximación propuesta con la de Box, en una distribución parabólica convexa en $[-4, 4]$ (tamaño de muestra $n = 5$).

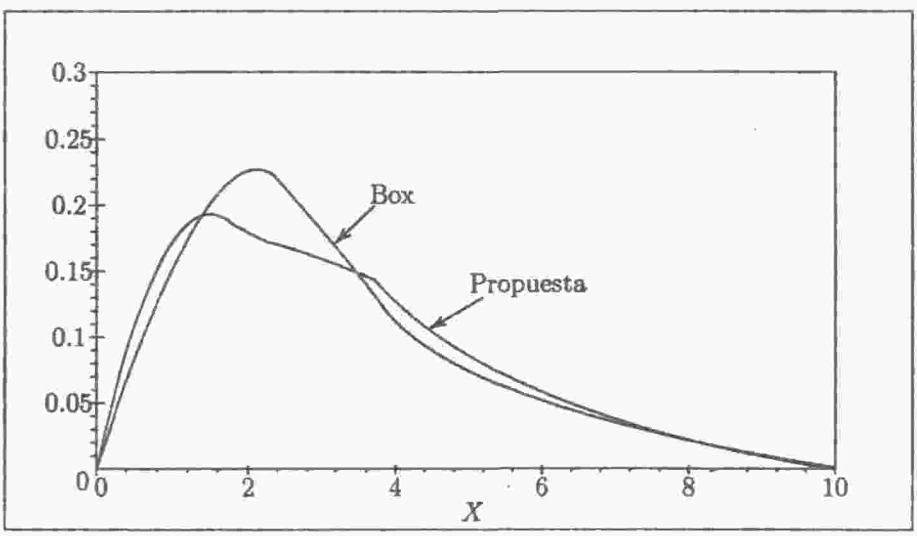


Tabla 1: Distribución uniforme

TAMAÑO DE	APROXIMACION DE BOX					
MUESTRA	β	α	MEDIA	VAR.	MOM. 3	MOM. 4
5	1.39	3.84	5.3376	7.41926	20.62555	251.14499
20	0.2414	22.09	5.332526	1.28727	0.62149	5.42129
50	0.08968	59.47	5.3332696	0.47828	0.08578	0.70935

APROXIMACION PROPUESTA							
$\alpha 1$	$\beta 1$	$\alpha 2$	$\beta 1$	MEDIA	VAR	MOM. 3	MOM. 4
3.2549	1.0803	11.4829	0.6217	5.328	7.399	8.6	151.74
15.8214	0.304	37.7126	0.1441	5.12	1.22	0.239	4.83
59.6688	0.085	81.4621	0.0686	5.33	0.474	0.044	0.671

MOMENTOS TEORICOS				SUMA DE
MEDIA	VAR.	MOM. 3	MOM. 4	CUADRADOS
5.3333	7.4	8.6	151.7	1.10E-05
5.3333	1.29	0.21	4.77	0.053
5.3333	0.48	0.046	0.67	3.71E-05

Valores de 1°, 2°, 3° y 4° momentos para la distribución de S^2 en una población uniforme.

Tabla 2: Distribución parabólica convexa

TAMAÑO DE	APROXIMACION DE BOX					
MUESTRA	β	α	MEDIA	VAR.	MOM. 3	MOM. 4
5	1.0514	3.0435	3.199935	3.366441	7.07468	56.27279
20	0.199699	16.024	3.199976	0.639032	0.25522	1.37799
50	0.075755	42.2414	3.199994	0.242415	0.03672	0.18464

APROXIMACION PROPUESTA							
$\alpha 1$	$\beta 1$	$\alpha 2$	$\beta 1$	MEDIA	VAR	MOM. 3	MOM. 4
2.9557	0.6862	7.9352	0.5676	3.266	3.507	4.569	40.1
12.5329	0.2408	25.2904	0.1331	3.19	0.618	0.162	1.271
36.5953	0.08488	54.4418	0.06043	3.198	0.24	0.0255	0.179

MOMENTOS TEORICOS				SUMA DE
MEDIA	VAR.	MOM. 3	MOM. 4	CUADRADOS
3.2	3.37	4.5	40.12	2.80E-02
3.2	0.64	0.15	1.26	8.30E-04
3.2	0.24	0.0261	0.18	5.50E-07

Valores de 1°, 2°, 3° y 4° momentos para la distribución de S^2 en una población parabólica convexa.

Tabla 3: Distribución triangular superior pura

TAMAÑO DE	APROXIMACION DE BOX					
MUESTRA	β	α	MEDIA	VAR.	MOM. 3	MOM. 4
5	1.01333	2.638	2.66708456	2.702636	5.47732	38.56375
20	0.2	13.287	2.6574	0.53148	0.212590	0.97496
50	0.076838	34.70528	2.666684305	0.204902	0.03148	0.13321

APROXIMACION PROPUESTA							
$\alpha 1$	$\beta 1$	$\alpha 2$	$\beta 1$	MEDIA	VAR	MOM. 3	MOM. 4
3.2433	0.437	6.1425	0.6141	2.687	2.732	4.297	29.037
11.1032	0.2242	18.3089	0.1546	2.66	0.527	0.162	0.919
33.4476	0.07414	50.3106	0.05686	2.67	0.21	0.0168	0.131

MOMENTOS TEORICOS				SUMA DE
MEDIA	VAR.	MOM. 3	MOM. 4	CUADRADOS
2.67	2.7	4.27	29.04	2.00E-03
2.67	0.54	0.14	0.91	8.30E-04
2.67	0.21	0.0017	0.13	7.23E-07

Valores de 1°, 2°, 3° y 4° momentos para la distribución de S^2 en una población triangular superior pura.