SIMULACION EN EL COMPUTADOR DEL TEOREMA CENTRAL DEL LIMITE

Oscar Brand V.
Departamento de Matemática
Universidad del Valle

INTRODUCCION

El concepto de la Distribución de las Medias Muestrales, es sin lugar a dudas, el desarrollo de mayor trascendencia e importancia de la estadística clásica. Lo anterior es corroborado en todos los textos de estadística, en los cuales el desarrollo en cuestion, es descrito como el *Teorema Central del Limite.*

Sinembargo, y a pesar de que el teorema es fundamental en la teoría de la inferencia, el concepto resulta díficil de entender y ser asimilado, aún por estudiantes de pregrado en estadística, no obstante el énfasis que autores y profesores hacen con el propósito de procurar una buena asimilación y comprensión del tema, mediante la solución de ejercicios sobre los cuales en la mayor parte de los casos se cae en una simple mecanización.

En un esfuerzo por lograr una asimilación del teorema, en algunos textos se propone como ejercicio la toma sucesiva de muestras aleatorias del mismo tamaño y con reemplazamiento a partir de una población finita, en donde una vez estimadas las medias muestrales se procede a graficar su histograma con el fín de observar la forma de su distribución etc.

Aunque el propósito del ejercicio es bueno, dicho ejercicio resulta sumamente tedioso debido a que en su ejecución se consume demasiado tiempo. Es justamente ésta la razón principal por la cual resulta de gran valor pedagógico el poder realizar dicha *simulación* con la ayuda del *computador*, pues en un tiempo relativamente muy corto, se pueden simular muchas muestras del mismo tamaño obteniendo, a partir de cada muestra las estimaciones básicas de los parámetros *de la población* entre ellos las de la media poblacional cuya distribución e histograma se pueden graficar.

Esta simulación la han realizado estudiantes de los cursos de pregrado en Estadística I y II a mi cargo y el resultado observado ha sido satisfactorio desde el punto de vista de su propósito, ya que ha permitido al estudiante no solamente comprobar y asimilar rápidamente el concepto, sino que tambien ha ayudado a generar ese sentimiento de credibilidad sobre la veracidad de los conceptos teóricos y a valorar la utilidad y las bondades de la estadística.

En éste articulo, se presenta además de una breve reseña histórica, un enunciado clásico del teorema, una descripción de los resultados y en la parte final o apéndice, una presentación a manera de ejemplo de los resultados obtenidos con la simulación utilizando uno de los archivos previamente grabados el cual contiene una 'población' de 400 datos.

ORIGEN DEL TEOREMA CENTRAL DEL LIMITE

Su origen se remonta al año 1733 cuando DeMoivre probó el teorema para variables aleatorias Bernoulli donde resulta una variable binomial. Una extensión de este teorema fué planteada posteriormente en 1980 por Laplace y Gauss, pero en revisiones posteriores se plantearon condiciones satisfactorias y exactas para su validez las cuales fueron definidas por Lyapunov en 1901 constituyéndose, desde entonces, en un problema sobresaliente de la teoría de la probabilidad desde sus comienzos hasta la decada de 1930.

En 1920 G. Polya le dió a este problema el nombre del *Teorema Central del Límite de la Teoría de la Probabilidad* aunque otros aseguran que un nombre más adecuado sería "Teorema de la Convergencia Normal" (Normal convergente Theorem).

En 1920 y 1930 el método de la función característica se usó para extender el teorema en varias direcciones y para obtener las condiciones necesarias y suficientes para su validez, en el caso en el cual las variables aleatorias son independientes. Aún se continua realizando un intenso trabajo para extender el *Teorema Central del Límite* para el caso de variables aleatorias dependientes.

Aunque existen muchas formas de expresar matemáticamente el teorema, es conveniente manifestar que es muy común encontrar en los textos de Estadística Aplicada, una forma incorrecta y vaga de los enunciados al teorema para poblaciones finitas.

Desde el punto de vista de las aplicaciones estadísticas se consideran básicamente dos versiones del *Teorema Central del Límite* y su validez es demostrada matemáticamente. Estas dos condiciones se enuncian a continuación:

(1) Si las variables aleatorias $X_1, X_2, \cdots$ son independientes e idénticamente distribuidas.

(2) Si las variables aleatorias $X_1, X_2, \cdots$ son independientes pero no idénticamente distribuidas.

Sinembargo, la forma clásica de enunciar este teorema es la siguiente:
Si $X_1, X_2, \cdots, X_n$ *denotan a los elementos de una muestra aleatoria de tamaño n, tomados de una distribución la cual tiene media* μ *y varianza positiva* σ^2, *entonces la variable aleatoria* $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ *tiene una distribución límite la cual es Normal con media cero y varianza uno.*

PROGRAMA DE SIMULACION DEL TEOREMA CENTRAL DEL LIMITE

El programa de simulación del *Teorema Central del Límite o Distribución de las Medias Muestrales* (TCL), está escrito en lenguaje Basic y puede operar en cualquier ambiente con sistema operacional D.O.S. y tarjeta graficadora.

Para su ejecución solo se requiere una vez introducido el diskette en el drive, teclear **tcl <enter>** y el programa que es iterativo, permitirá al estudiante con sólo teclear las respuestas a los interrogantes que se presentan, proceder con la simulación para lo cual es necesario empezar considerando una Población finita de elementos como resultado de un proceso aletorio, por ejemplo, el espacio muestral que se origina mediante el lanzamiento de una moneda o un dado, cuyos datos se pueden entrar durante el ejercicio o considerar otro proceso cuyos datos pueden leerse desde un archivo previamente grabado, obteniendo:

1. El valor de los parámetros de la población y el histograma de la distribución de la población.

2. Muestras aleatorias (no más de 800) de tamaño no superior a 900, dónde a partir de cada muestra se estima: La media, la varianza, el error estándar (desv. estándar de la media) y el intervalo de confianza para la media poblacional μ con una descripción sobre si el verdadero valor de μ está o no en el intervalo.

3. Información sobre el número de intervalos de confianza que contienen a la verdadera media poblacional μ e información sobre el número de intervalos de confianza que fallan en el contenido de la media por razones de la aleatoriedad.

4. Estimación de la media de las medias muestrales, su comparación con la media poblacional o parámetro μ , estimación de la desviación estándar de las medias o *error estándar* y su comparación con el parámetro poblacional.

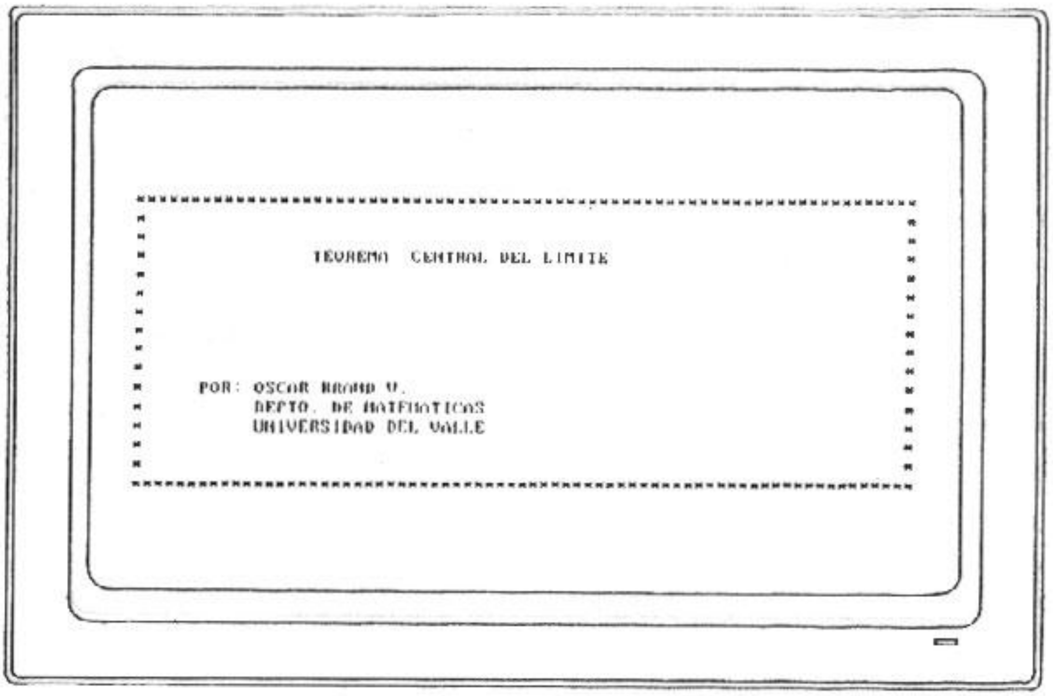
$$\sigma_{\bar{x}} = \sigma / \sqrt{n}$$

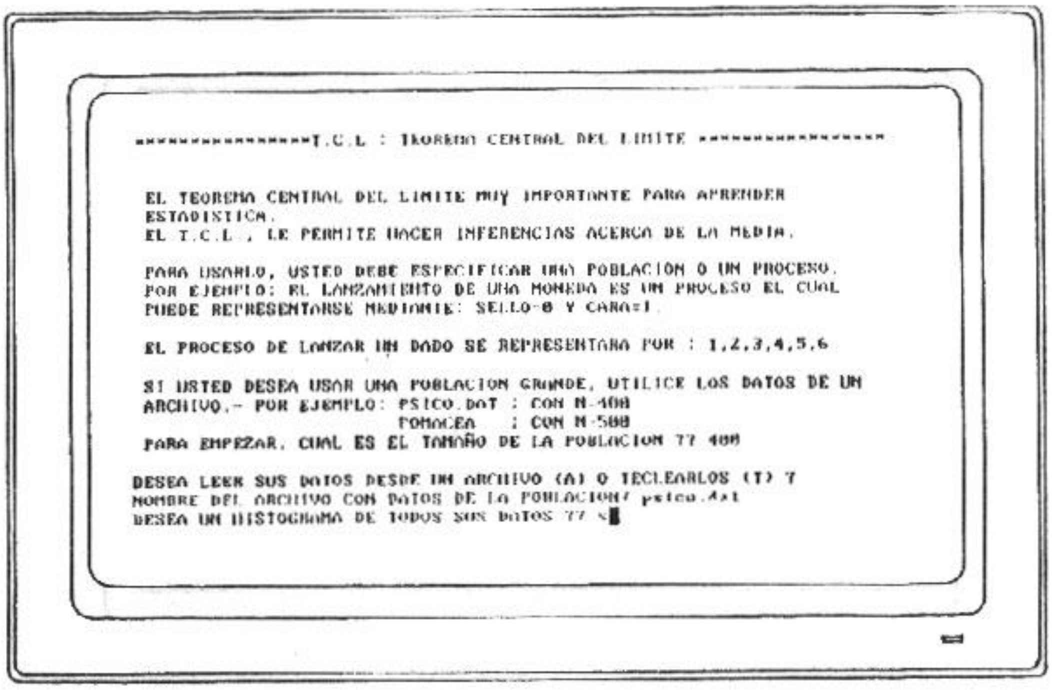
5. Histograma de la distribución de las medias y la superposición de la Distribución Normal definida con los parámetros de la población, lo que permite observar si el histograma se asemeja a la Distribución Normal. En caso contrario, se correrán nuevas simulaciones definiendo tamaños de muestras mayores, lo que permitirá observar cómo la aproximación se dá a medida que se aumenta el tamaño de la muestra.

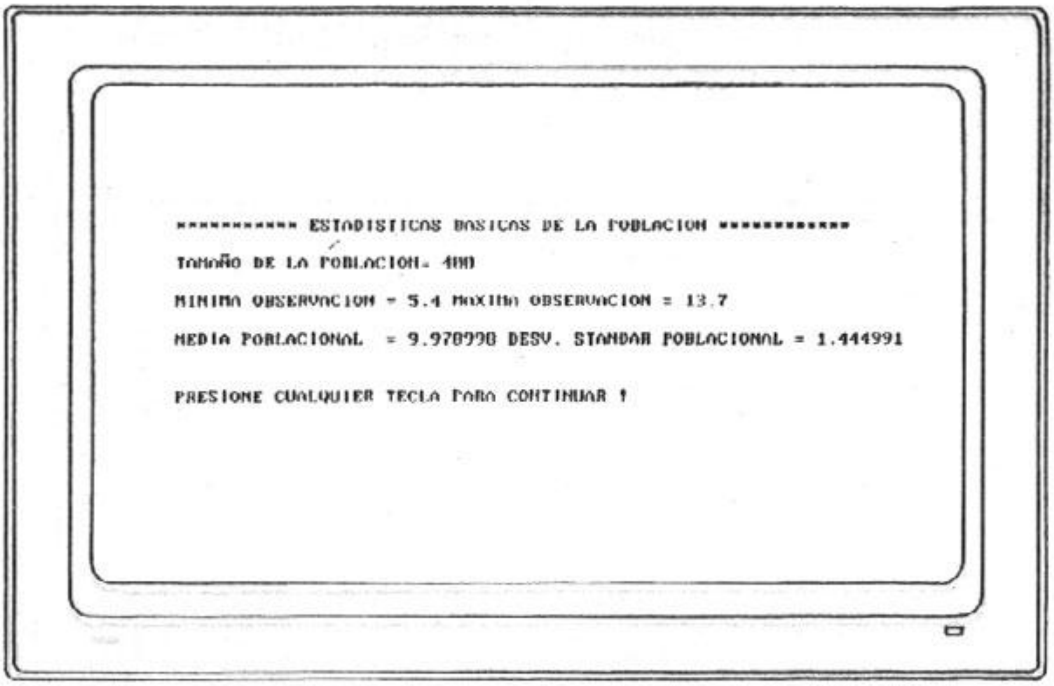
6. Finalmente, se refuerza el hecho de que en la práctica sólo es necesario la obtención de una muestra, procediéndose a su selección y estimación de las estadísticas básicas e histograma.

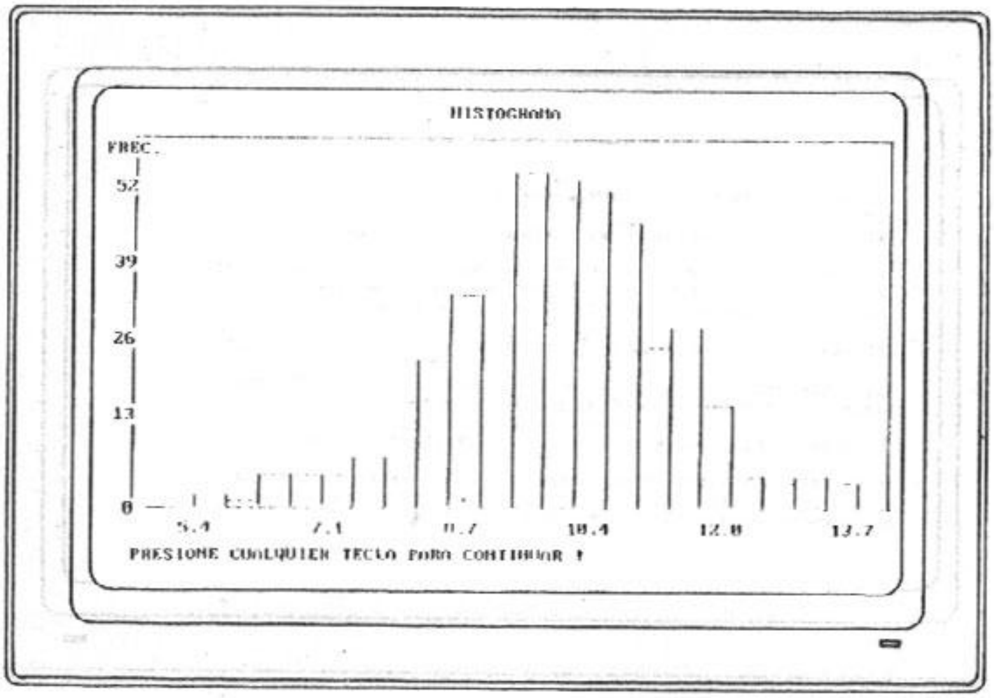
Esta simulación puede ser repetida por el estudiante todas las veces que desee, utilizando además diferentes poblaciones finitas, hasta lograr su completa asimilación.

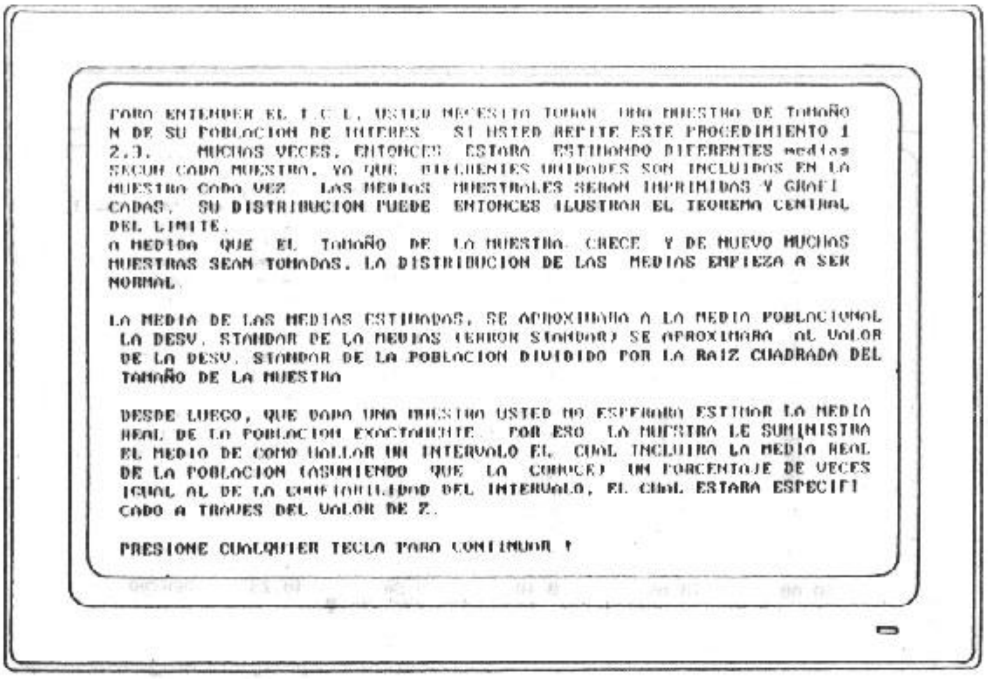
Una copia de este programa puede solicitarse al Centro de Cálculo del Departamento de Matemáticas de la Universidad del Valle, proporcionando el diskette para su copia e indicando el tipo de tarjeta graficadora que utiliza su computador.

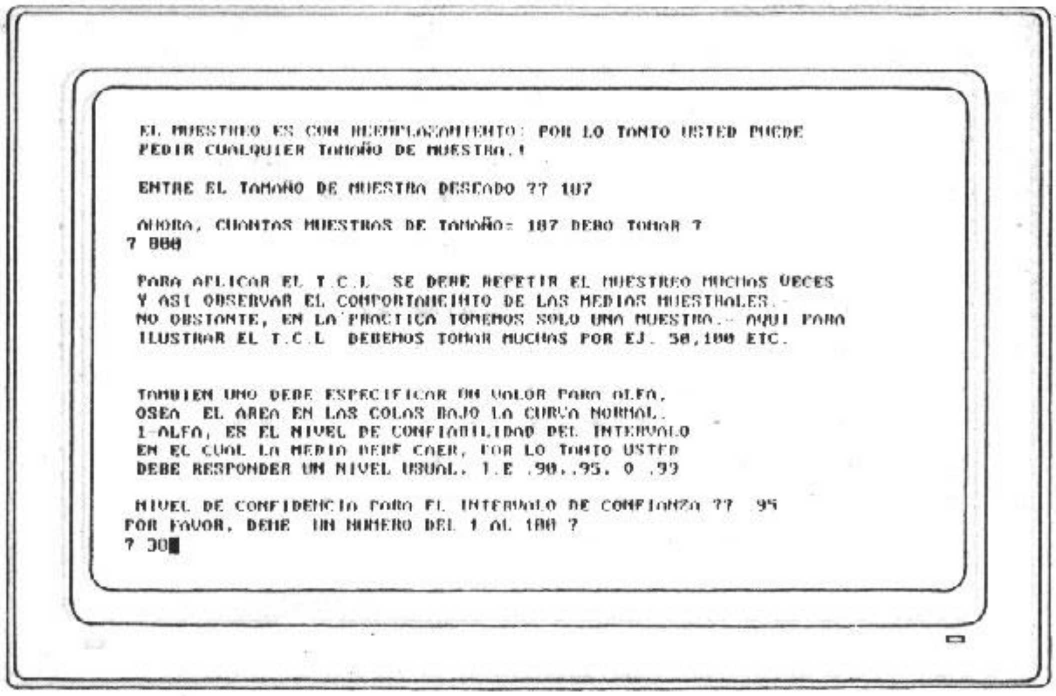

TEOREMA CENTRAL DEL LIMITE
POR: OSCAR BRAND U.
DEPTO. DE MATEMATICAS
UNIVERSIDAD DEL VALLE

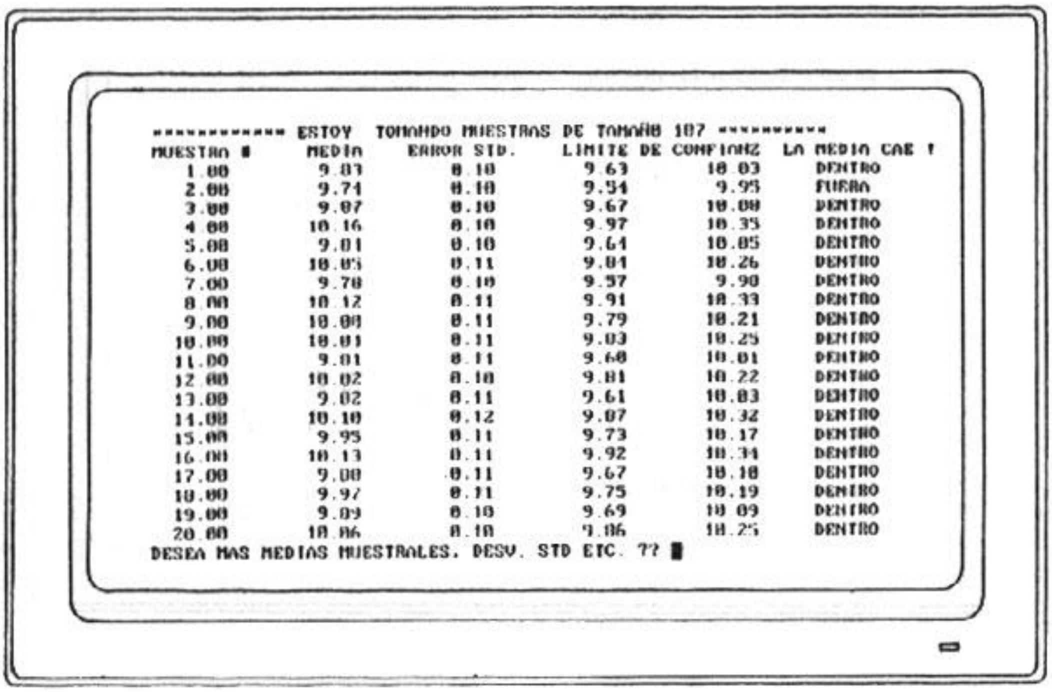

T.C.L : TEOREMA CENTRAL DEL LIMITE
EL TEOREMA CENTRAL DEL LIMITE MUY IMPORTANTE PARA APRENDER ESTADISTICA.
EL T.C.L., LE PERMITE HACER INFERENCIAS ACERCA DE LA MEDIA.
PARA USARLO, USTED DEBE ESPECIFICAR UNA POBLACION O UN PROCESO.
POR EJEMPLO: EL LANZAMIENTO DE UNA MONEDA ES UN PROCESO EL CUAL PUEDE REPRESENTARSE MEDIANTE: SELLO=0 Y CARA=1.
EL PROCESO DE LANZAR UN DADO SE REPRESENTARA POR : 1,2,3,4,5,6
SI USTED DESEA USAR UNA POBLACION GRANDE, UTILICE LOS DATOS DE UN ARCHIVO.- POR EJEMPLO: PSICO.DAT : CON N=400
POMACEA : CON N=500
PARA EMPEZAR, CUAL ES EL TAMAÑO DE LA POBLACION ?? 400
DESEA LEER SUS DATOS DESDE UN ARCHIVO (A) O TECLEARLOS (T) ?
NOMBRE DEL ARCHIVO CON DATOS DE LA POBLACION? psico.dat
DESEA UN HISTOGRAMA DE TODOS SUS DATOS ?? s


*********** ESTADISTICAS BASICAS DE LA POBLACION ***********
TAMAÑO DE LA POBLACION= 400
MINIMA OBSERVACION = 5.4 MAXIMA OBSERVACION = 13.7
MEDIA POBLACIONAL = 9.978998 DESV. STANDAR POBLACIONAL = 1.444991
PRESIONE CUALQUIER TECLA PARA CONTINUAR !


HISTOGRAMA
FREC.
52
39
26
13
0
5.4
7.1
8.7
10.4
12.0
13.7
PRESIONE CUALQUIER TECLA PARA CONTINUAR !


PARA ENTENDER EL T.C.L, USTED NECESITA TOMAR UNA MUESTRA DE TAMAÑO
N DE SU POBLACION DE INTERES. SI USTED REPITE ESTE PROCEDIMIENTO 1
2.3. MUCHAS VECES, ENTONCES ESTARA ESTIMANDO DIFERENTES medias
SEGUN CADA MUESTRA, YA QUE DIFERENTES UNIDADES SON INCLUIDAS EN LA
MUESTRA CADA VEZ. LAS MEDIAS MUESTRALES SERAN IMPRIMIDAS Y GRAFI
CADAS. SU DISTRIBUCION PUEDE ENTONCES ILUSTRAR EL TEOREMA CENTRAL
DEL LIMITE.
A MEDIDA QUE EL TAMAÑO DE LA MUESTRA CRECE Y DE NUEVO MUCHAS
MUESTRAS SEAN TOMADAS. LA DISTRIBUCION DE LAS MEDIAS EMPIEZA A SER
NORMAL.
LA MEDIA DE LAS MEDIAS ESTIMADAS, SE APROXIMARA A LA MEDIA POBLACIONAL
LA DESV. STANDAR DE LA MEDIAS (ERROR STANDAR) SE APROXIMARA AL VALOR
DE LA DESV. STANDAR DE LA POBLACION DIVIDIDO POR LA RAIZ CUADRADA DEL
TAMAÑO DE LA MUESTRA
DESDE LUEGO, QUE DADA UNA MUESTRA USTED NO ESPERARA ESTIMAR LA MEDIA
REAL DE LA POBLACION EXACTAMENTE. POR ESO LA MUESTRA LE SUMINISTRA
EL MEDIO DE COMO HALLAR UN INTERVALO EL CUAL INCLUIRA LA MEDIA REAL
DE LA POBLACION (ASUMIENDO QUE LA CONOCE) UN PORCENTAJE DE VECES
IGUAL AL DE LA CONFIABILIDAD DEL INTERVALO, EL CUAL ESTARA ESPECIFI
CADO A TRAVES DEL VALOR DE Z.
PRESIONE CUALQUIER TECLA PARA CONTINUAR !

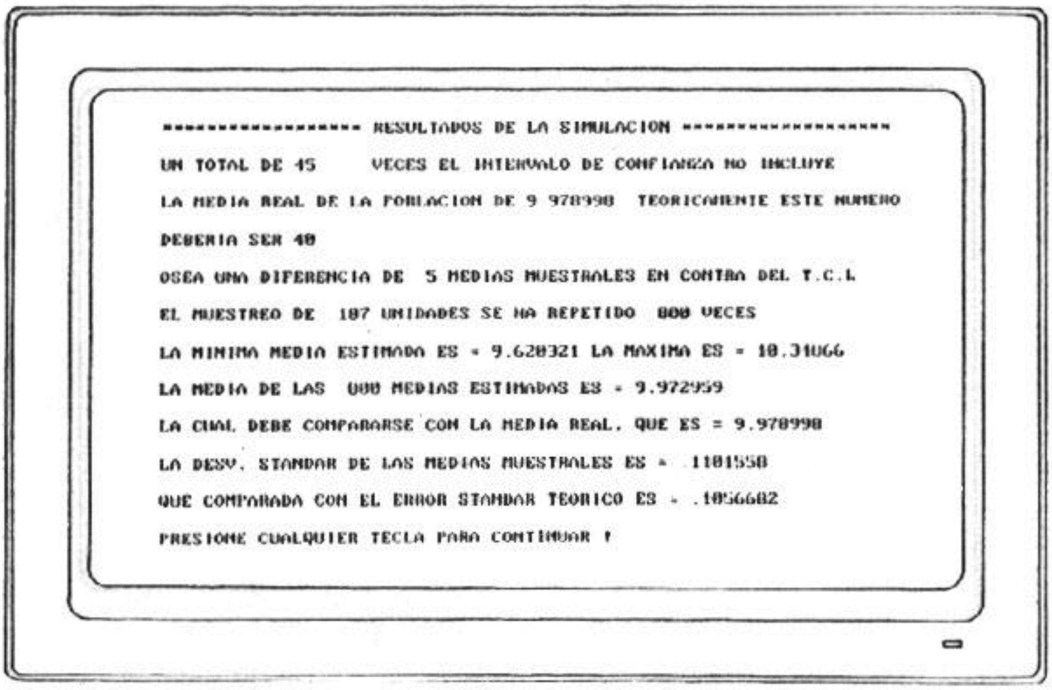

EL MUESTREO ES CON REEMPLAZAMIENTO: POR LO TANTO USTED PUEDE
PEDIR CUALQUIER TAMAÑO DE MUESTRA.!
ENTRE EL TAMAÑO DE MUESTRA DESEADO ?? 107
AHORA, CUANTAS MUESTRAS DE TAMAÑO= 107 DEBO TOMAR ?
? 800
PARA APLICAR EL T.C.L SE DEBE REPETIR EL MUESTREO MUCHAS VECES
Y ASI OBSERVAR EL COMPORTAMIENTO DE LAS MEDIAS MUESTRALES.-
NO OBSTANTE, EN LA PRACTICA TOMEMOS SOLO UNA MUESTRA.- AQUI PARA
ILUSTRAR EL T.C.L DEBEMOS TOMAR MUCHAS POR EJ. 50,100 ETC.
TAMBIEN UNO DEBE ESPECIFICAR UN VALOR PARA ALFA,
OSEA EL AREA EN LAS COLAS BAJO LA CURVA NORMAL.
1-ALFA, ES EL NIVEL DE CONFIABILIDAD DEL INTERVALO
EN EL CUAL LA MEDIA DEBE CAER, POR LO TANTO USTED
DEBE RESPONDER UN NIVEL USUAL. I.E .90,.95, O .99
NIVEL DE CONFIDENCIA PARA EL INTERVALO DE CONFIANZA ?? 95
POR FAVOR, DEME UN NUMERO DEL 1 AL 100 ?
? 30



************ ESTOY TOMANDO MUESTRAS DE TAMAÑO 107 **********

MUESTRA #	MEDIA	ERROR STD.	LIMITE DE CONFIANZ		LA MEDIA CAE !
1.00	9.83	0.10	9.63	10.03	DENTRO
2.00	9.74	0.10	9.54	9.95	FUERA
3.00	9.87	0.10	9.67	10.08	DENTRO
4.00	10.16	0.10	9.97	10.35	DENTRO
5.00	9.81	0.10	9.64	10.05	DENTRO
6.00	10.05	0.11	9.84	10.26	DENTRO
7.00	9.78	0.10	9.57	9.90	DENTRO
8.00	10.12	0.11	9.91	10.33	DENTRO
9.00	10.00	0.11	9.79	10.21	DENTRO
10.00	10.01	0.11	9.83	10.25	DENTRO
11.00	9.81	0.11	9.60	10.01	DENTRO
12.00	10.02	0.10	9.81	10.22	DENTRO
13.00	9.82	0.11	9.61	10.03	DENTRO
14.00	10.10	0.12	9.87	10.32	DENTRO
15.00	9.95	0.11	9.73	10.17	DENTRO
16.00	10.13	0.11	9.92	10.34	DENTRO
17.00	9.88	0.11	9.67	10.10	DENTRO
18.00	9.97	0.11	9.75	10.19	DENTRO
19.00	9.89	0.10	9.69	10.09	DENTRO
20.00	10.06	0.10	9.86	10.25	DENTRO

DESEA MAS MEDIAS MUESTRALES. DESV. STD ETC. ??



***************** RESULTADOS DE LA SIMULACION ******************

UN TOTAL DE 45 VECES EL INTERVALO DE CONFIANZA NO INCLUYE

LA MEDIA REAL DE LA POBLACION DE 9.978998 TEORICAMENTE ESTE NUMERO

DEBERIA SER 40

OSEA UNA DIFERENCIA DE 5 MEDIAS MUESTRALES EN CONTRA DEL T.C.L

EL MUESTREO DE 107 UNIDADES SE HA REPETIDO 800 VECES

LA MINIMA MEDIA ESTIMADA ES = 9.620321 LA MAXIMA ES = 10.31066

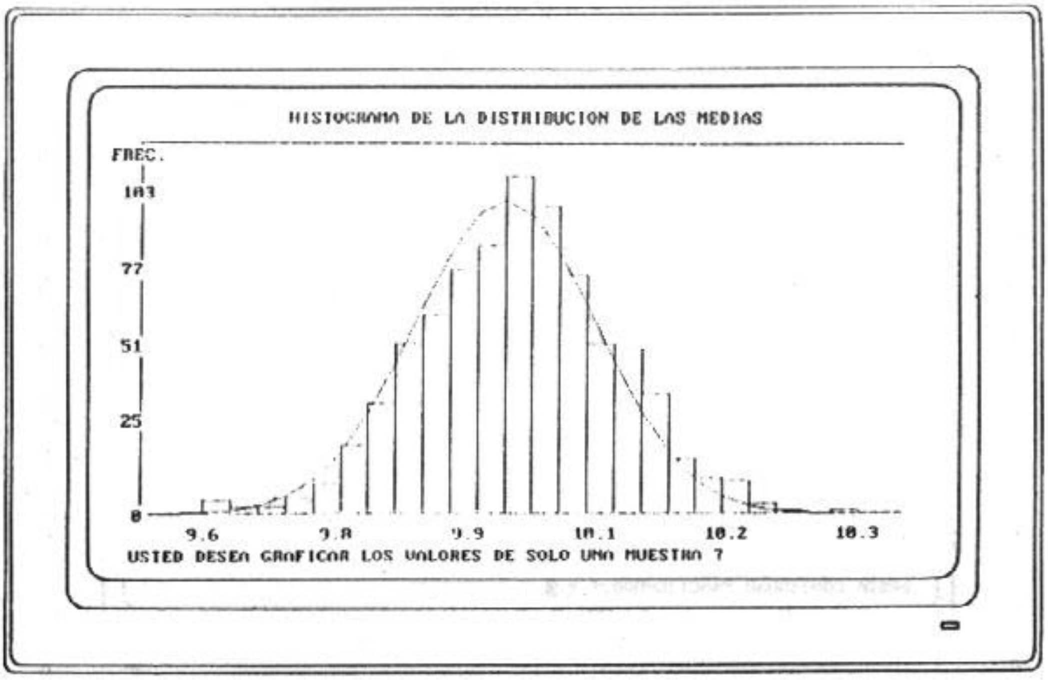
LA MEDIA DE LAS 800 MEDIAS ESTIMADAS ES = 9.972959

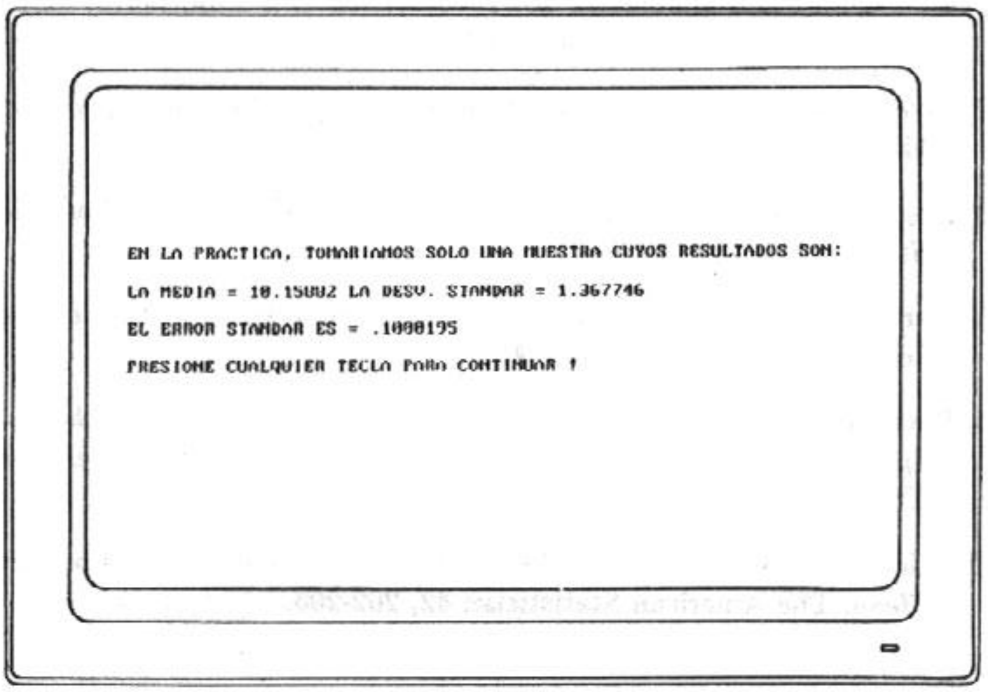
LA CUAL DEBE COMPARARSE CON LA MEDIA REAL, QUE ES = 9.978998

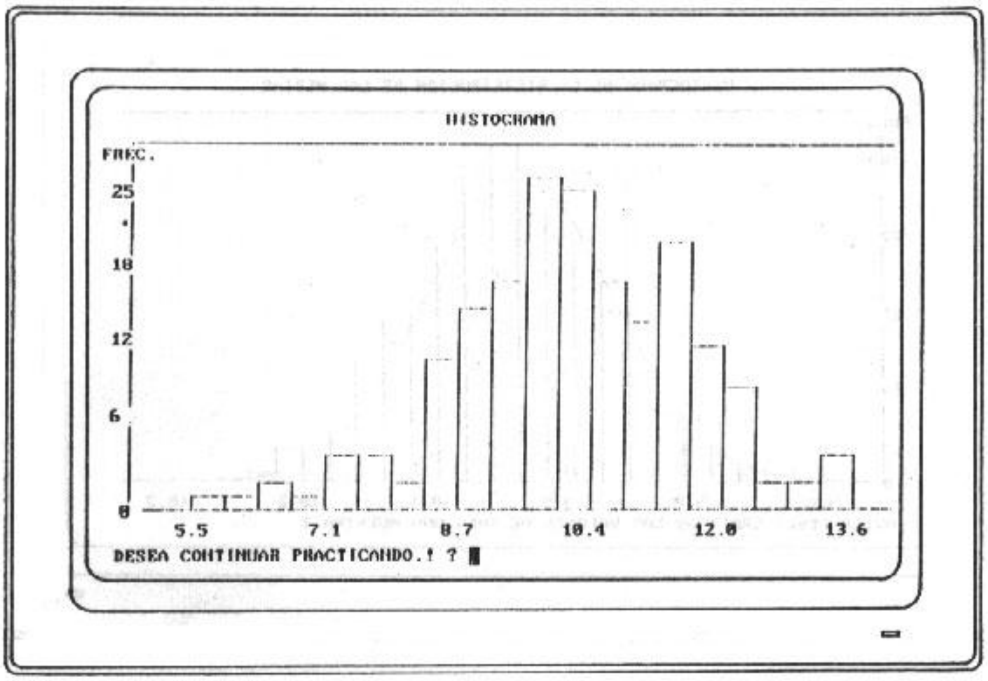
LA DESV. STANDAR DE LAS MEDIAS MUESTRALES ES = .1101558

QUE COMPARADA CON EL ERROR STANDAR TEORICO ES = .1056682

PRESIONE CUALQUIER TECLA PARA CONTINUAR !


HISTOGRAMA DE LA DISTRIBUCION DE LAS MEDIAS
FREC.
103
77
51
25
0
9.6
9.8
9.9
10.1
10.2
10.3
USTED DESEA GRAFICAR LOS VALORES DE SOLO UNA MUESTRA ?


EN LA PRACTICA, TOMARIAMOS SOLO UNA MUESTRA CUYOS RESULTADOS SON:
LA MEDIA = 10.15002 LA DESV. STANDAR = 1.367746
EL ERROR STANDAR ES = .1000195
PRESIONE CUALQUIER TECLA PARA CONTINUAR !



Bibliografía

[1] Barret J.P. and Goldsmith L. (1976) *When is n suffucienthy large?*, The American Statistician 30, 67-70.

[2] Dixon W.J. y Massey F.J. *Introducción al Análisis Estadístico*, McGraw-hill, 1965.

[3] Hogg W. Robert and Craig T. Allen., *Introduction to Mathematical Statististics*, The Maccmillan Company, 1970.

[4] Parzen Emanuel., *Modern Probability Theory and Its Applications*, John Wiley and Sons, Inc. 1970.

[5] Plane R. Donald and Gordon K.R. (1982) *A Simple Proof of the Non-applicability of the Central Limit Theorem to Finite Populations*, The American Statistician 36, 175-176.

[6] Shie Shien Yang (1988) *A Central Limit Theorem for the Bootstrap Mean*, The American Statistician 42, 202-203.