# El Problema de los Taxis y el Problema del Circulante.

Ricardo Miró
Consejo de la Magistratura de la Nación
Área de Procesamiento de Datos

## 1. Introducción.

Como es sabido, varias aplicaciones tecnológicas de gran importancia surgieron o fueron puestas a punto durante la Segunda Guerra Mundial. No parece oportuno efectuar una lista completa de estos aportes, pero para tener una idea de los mismos, basta nombrar a la Informática, a la Tecnología de las Microondas y a la Investigación Operativa.

A propósito de la Tecnología de las Microondas, conviene recordar que de ella deriva el radar. Hacia el bienio 1942/43, antes de que ese instrumento crucial fuera completamente desarrollado, la defensa aérea inglesa estaba muy exigida. Para prevenir a la población civil de los ataques aéreos alemanes, contaba únicamente con una cadena de globos cautivos en concordancia con una red bastante amplia de reflectores de arco voltaico. Se tornaba obligatorio, entonces, disminuir en todo lo posible la capacidad de producción de las fábricas aeronáuticas del Tercer Reich. Ahora bien: no es lo mismo destruir un taller que tiene una capacidad de gestar 20 planeadores deportivos de lujo anuales, que un complejo bélico capaz de producir 500 máquinas de guerra por mes. Mediante una organizada red de información, la artillería antiaérea fue instruida por el Alto Mando para tomar nota del número de serie de todos los aparatos abatidos. Con este sencillo método, y mediante un procesamiento de datos posterior de iguales características, fue posible tener una estimación muy aceptable de la cantidad de aviones de combate en pie de guerra, y por lo tanto, de la capacidad de producción enemiga, [3].

Este artículo mostrará de qué manera es posible estimar el número N de individuos de una población dada, en donde cada uno de sus integrantes se identifica por un número natural único, ubicado en el intervalo [1, N]. Le estimación de N se realiza por muestreo, tal como se procederá a continuación.

Por obvias connotaciones, la literatura específica discurre sobre este tema llamándolo “el problema de los taxis” [6]. Como los billetes que constituyen el dinero circulante de un país tienen un número de serie con análogas características, nada impide concebirlo también como “el problema del circulante”.

## 2. Una Característica Algebraica de las Poblaciones Numeradas.

Antes de entrar de lleno en el problema que motiva el presente artículo, tal vez convenga observar cómo quedan caracterizados dos estadísticos importantes - el *promedio*, y la *varianza* -, en una población cuyos individuos son todos los números naturales entre 1 y *N*.

Lo que se desea hacer, luego, es determinar el promedio y la varianza de una variable aleatoria con la capacidad de producir todos los números de licencias de taxis en una gran ciudad como Buenos Aires.

Las consideraciones que siguen son teóricas, pero tienen mucha importancia cuando se aborden posteriormente el problema de la estimación del número natural *N*, que tal como se ha mencionado, designará la cantidad total de licencias de taxis. Se supondrá, para abreviar la discusión, que todas las licencias de 1 hasta *N* están habilitadas para operar y por lo tanto, pueden eventualmente llegar a verse en un relevamiento realizado ad hoc.

Sea entonces *L* el conjunto de todas las licencias de taxis existentes. Formalmente se puede escribir:

$$L = \{ k \in N / 1 \leq k \leq N \}$$

Resultará muy fácil determinar el promedio de la población numérica *L*. Se lo designará como es habitual con la letra $\mu$ . Luego:

$$\mu = \frac{1}{N} \sum_{k=1}^{N} k \qquad (1)$$

Ahora bien: la suma que aparece en el segundo miembro de la igualdad (1) es una clarísima progresión aritmética, en donde

$$\sum_{k=1}^{N} k = \frac{N(N+1)}{2} \qquad (2)$$

Entonces, reemplazando (2) en (1) quedará:

$$\mu = \frac{N+1}{2} \qquad (3)$$

O lo que es lo mismo:

$$N = 2\mu - 1 \qquad (4)$$

Finalmente, la ecuación (4) establece una sencilla vinculación entre el promedio $\mu$ y la cantidad de licencias *N*.

Determinemos ahora la varianza de la variable aleatoria tratada. Sin generar ninguna confusión llamemos también *L* a la variable aleatoria discreta , que produce números de licencia $k\ (1 \leq k \leq N)$, cada uno de ellos con igual probabilidad $\frac{1}{N}$, entonces [2] se sabe que

$$Var(L) = \sum_{k=1}^{N} \left( k - \frac{N+1}{2} \right)^2 \cdot \frac{1}{N} \qquad (5)$$

Esta expresión, hay que reconocerlo, es un poco más trabajosa que el promedio. De todas maneras, es fácilmente simplificable. Desarrollando el binomio que aparece en (5), se tiene que

$$Var(L) = \frac{1}{N}\left[\sum_{k=1}^{N} k^2 + N\left(\frac{N+1}{2}\right)^2 - \frac{N\left(N+1\right)^2}{2}\right] \qquad (6)$$

En (6), aparece la suma de los primeros *N* cuadrados. Se demuestra por inducción que

$$\sum_{k=1}^{N} k^2 = \frac{1}{3}\left(N^3 + \frac{3N^2}{2} + \frac{N}{2}\right)$$

Reemplazando esto último en (6), sale finalmente que

$$Var(L) = \frac{N^2 - 1}{12}. \qquad (7)$$

## 3. Comportamiento de la Varianza Muestral.

La expresión citada en (7) es de utilidad. Tal como se verá más adelante, la varianza poblacional interviene en la *confiabilidad* con la que se estima la cantidad desconocida *N*. Pero como la varianza es asimismo desconocida, resulta necesaria estimarla por muestreo, y esto es fácil de realizar.

Mediante un sencillo programa para PC, es posible simular la generación de unas 200 varianzas muestrales, cuyos tamaños toman todos los valores de 1 a 200 licencias, para ver que es lo que sucede. El gráfico siguiente ilustra el resultado obtenido, suponiendo N= 40000.

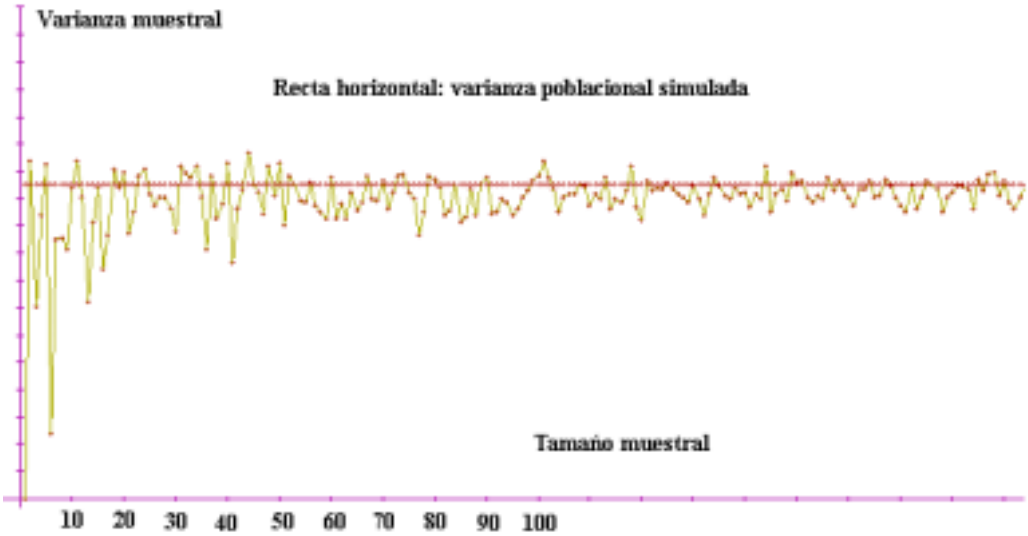


Fig. 1 : Estabilización de la varianza muestral.

Se observa con toda claridad que la varianza muestral comienza a oscilar alrededor del verdadero valor (representado por la recta horizontal de ordenada igual a 11.547), prácticamente a partir de las muestras que tienen 30 o más licencias. Este comportamiento es absolutamente general, y puede demostrarse con todo rigor, tal como se hace en [6]. Por otra parte, la estimación de la varianza poblacional a través de una muestra $L_n$ de tamaño $n$ cuyo promedio muestral es $\overline{L_n}$, se realiza mediante la siguiente expresión:

$$Var_n(L) = \frac{1}{n-1}\sum_{k=1}^{n}\left(L_k - \mu_n\right)^2$$

con la cual se ha obtenido el gráfico de la Fig. 1.

## 4. Estimación de N a partir del Promedio Muestral.

La igualdad dada en (4), será utilizada ahora para estimar la cantidad de licencias totales $N$. Como no tenemos acceso a los registros municipales, podremos intentar estimar $\mu$ a partir de un muestreo que resulte significativo y a la vez fácil de hacer.

Se pueden anotar por ejemplo las primeras $n$ licencias que pasen por una esquina concurrida y calcular luego tranquilamente el promedio muestral $\mu_n$ de la manera obvia:

$$\mu_n = \frac{1}{n}\sum_{k=1}^{n} l_k \qquad (8)$$

Luego, en poder de este dato, proclamamos que

$$\overline{N} = 2 \cdot \mu_n \qquad (9)$$

es una buena estimación para la cantidad total $N$ de licencias de taxis. Esto, naturalmente, debe justificarse.

La primera observación que se puede hacer a la estimación (9) es que se ha sacrificado el "1" que aparece en la ecuación poblacional (4), dado el carácter grande de $N$.

La segunda observación es un poco más delicada y se refiere al comportamiento de los promedios muestrales. Cuando se toma una muestra considerablemente grande (de $n$ licencias o más) es de esperar que esa elección permita detectar licencias altas (muy cerca del $N$ poblacional) y licencias bajas, de tal manera que el promedio muestral $\mu_n$ no difiera ostensiblemente del promedio poblacional $\mu$. Es decir: es razonable suponer que cada licencia $k$ de todas las existentes, tenga la misma probabilidad de pasar ante nuestros ojos y quedar entonces debidamente registrada. Para ilustrar este último concepto, se pueden anotar unas 200 licencias y construir el histograma correspondiente, Los datos se capturaron en Bilinghurst y Corrientes, el sábado 25 de marzo del 2000 a las 11 de la mañana:

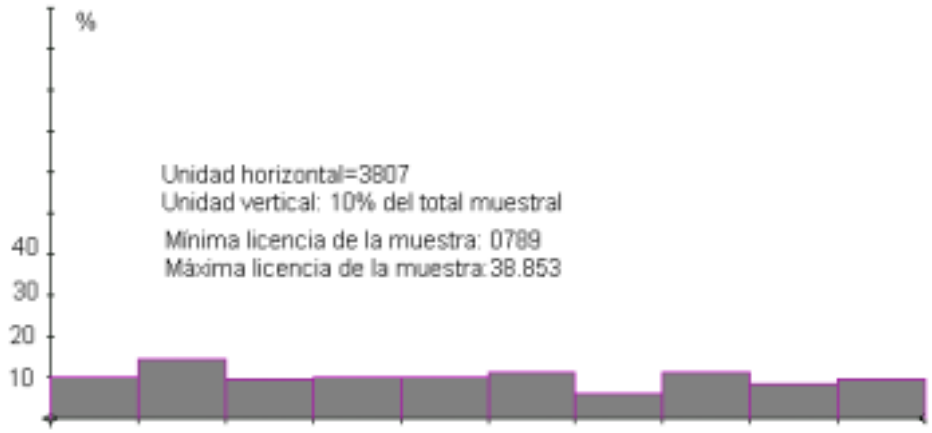


Fig. 2: Histograma de frecuencias de 200 licencias de taxis.

El histograma tiene diez divisiones de igual longitud. Se observa que ningún intervalo es privilegiado y cada uno de ellos tiene aproximadamente el 10% de las licencias de la muestra. Esto sugiere, ahora con mayor verosimilitud, que los números de licencias de taxis están uniformemente distribuidos. Luego, las consideraciones heurísticas hechas anteriormente sobre la composición muestral (cantidad de licencias altas y bajas) son correctas. Debe aclararse, sin embargo, que existen procedimientos muy específicos para inferir verosímilmente la función de distribución subyacente en un experimento como el de las licencias de taxis. Uno de ellos es el criterio de bondad de ajuste de Kolmogoroff-Smirnoff , que está muy tratado en las obras de referencia, [2], [6].

Hay un resultado vital en Estadística Matemática : el Teorema Central del Límite. Su demostración es bastante dura y está expuesta en muchos lugares con diferentes niveles de generalidad. Un enfoque muy accesible puede encontrarse en [2]. En esencia, este teorema establece que los promedios muestrales, sea cual fuera la distribución subyacente de la variable aleatoria original, se distribuyen de manera acampanada, donde esa campana tiende a confundirse con la campana normal a medida que aumenta el tamaño de la muestra.

Más adelante se ilustrará con una sencilla simulación en PC lo que se acaba de decir. El histograma de la Fig. 1 constituye el punto de partida para tal fin.

## 5. Añoranzas Estadísticas y el Teorema Central del Límite.

Recordemos la definición de valor medio $E(X)$ para una variable aleatoria discreta $X$, tal como la que estamos examinando ahora:

$$E(X) = \mu = \sum_{k=0}^{N} x_k \cdot p(x_k) \qquad (10)$$

( $p(x_k)$ es la probabilidad del suceso "$X = x_k$" )

Recordemos también la definición de varianza poblacional, para una variable aleatoria discreta de recorrido finito:

$$Var(X) = \sum_{k=1}^{N} (x_k - \mu)^2 \; p(x_k) = E\left(X - \mu\right)^2 \tag{11}$$

Con casi nada de esfuerzo, [4], se puede ver que

$$E(X+Y) = E(X) + E(Y) \tag{12}$$

Ahora bien: al desarrollar el miembro extremo derecho de (11), y usando (12) donde corresponda , se tiene de inmediato que la varianza de *X* se puede escribir como

$$Var(X) = EX^2 - E^2X \tag{13}$$

Y (13) provoca de inmediato que

i) $Var(aX) = a^2 \cdot Var(X)$, para cualquier constante *a*. (14)

ii) Si *a* es constante, *Var(a)*= 0.

Si $X$ e $Y$ son independientes, es decir, que se verifica que $E(XY) = E(X)E(Y)$, será:

iii) $Var(X+Y) = Var(X) + Var(Y)$ (15)

En efecto:

$$Var(X+Y) = E\left(X+Y\right)^2 - \left(E(X+Y)\right)^2 =$$

$$= EX^2 + EY^2 + E(XY) - E^2X - E^2Y - 2EX \cdot EY =$$

$$= Var\left(X\right) + Var(Y) + 2\left[E(XY) - E(X)E(Y)\right] =$$

$$= Var(X) + Var(Y)$$

Resulta apropiado preguntarse para qué sirve todo esto, y la respuesta es inmediata: sirve para entender una de las maneras utilizadas por la literatura para presentar el Teorema Central del Límite (TCL), [2], [6].

Sea ahora $Y$ una variable aleatoria cualquiera discreta de recorrido finito, con valor medio $\mu$ y varianza $Var(Y)$. Si por casualidad se tuviera que $\mu = 0$ y $Var(Y) = 1$, se dirá que la variable *Y* está normalizada o estandarizada: esto facilita la sistematización del estudio de muchos fenómenos estadísticos, reduciéndolos a casos de catálogo . En general, para una variable aleatoria arbitraria, *esto no sucede para nada, ni tiene tampoco por qué suceder*.

Pero entonces, con la artillería que nos proporcionan los resultados expuestos entre (10) y (15), debe ser muy fácil demostrar que la variable siguiente, de aspecto terrorífico:

$$Z = \frac{Y - E(Y)}{\sqrt{Var(Y)}}$$

tiene valor medio 1 y promedio 0 y por lo tanto siempre está normalizada.

Volvamos ahora a los taxis, para encarar la siguiente dramatización:

Frente a una esquina concurrida, el investigador anotará 20 licencias y las promediará. Toma a continuación otras 20 licencias y las promediará de nuevo. Repite esta operación unas doscientas veces más. Debe quedar claro, luego, que con este proceder estará generando *otra variable aleatoria* que se puede legítimamente llamar $\mu_{20}$, y que señala los promedios obtenidos al recoger varias veces 20 licencias que pasan por la esquina. En general, se puede considerar $\mu_n$ como la variable aleatoria que se genera anotando *n* licencias y promediándolas luego.

Calculemos ahora el valor medio y la varianza de esta variable aleatoria $\mu_n$.

Como se sabe que $\mu_n = \frac{\sum_{k=1}^{n} L_k}{n}$, entonces usando las propiedades anteriores - (10) a (15)- , sale de inmediato que $E(\mu_n) = \mu$. Es también muy fácil ver que $Var(\mu_n) = \frac{Var(L)}{n}$.

Entonces se obtiene la siguiente variable normalizada:

$$Z_L = \frac{\mu_n - E(\mu_n)}{\sqrt{Var(\mu_n)}} = \frac{\mu_n - \mu}{\sqrt{Var(L)}} \cdot \sqrt{n} \qquad (16)$$

El Teorema Central del Límite dice que *a medida que el tamaño muestral n crece, la función de distribución que gobierna $Z_L$ se parecerá más y más a la función de distribución normal*. Es decir: los histogramas construidos a partir de promedios de muestras de tamaño *n*, procesando los resultados obtenidos según (16), se ubicarán cada vez mejor debajo de la famosa campana de Gauss.

Este resultado muestra de manera muy explícita por qué la distribución normal es importante: la distribución inicial puede ser arbitraria – en el caso aquí tratado es uniforme- , pero la distribución de los promedios normalizados de muestras de longitud o tamaño *n* se *acampanan* más y más, hasta confundirse en el infinito con la campana de Gauss.

Tal vez se pueda decir que el TCL muestra cómo la Naturaleza, por sí sola, *tira al blanco* con los promedios muestrales. Y esta analogía no es caprichosa. La rosa de tiro generada por un tirador olímpico sobre la diana de papel sigue clásicamente una distribución normal radial, muy comprimida respecto del centro. La Naturaleza actúa como un tirador experto, al usar promedios normalizados generados sobre muestras de gran tamaño, que se comprimen en el “centro”, es decir el promedio 0 de la distribución normal. La siguiente serie de histogramas se obtuvo mediante una simulación hecha en PC.

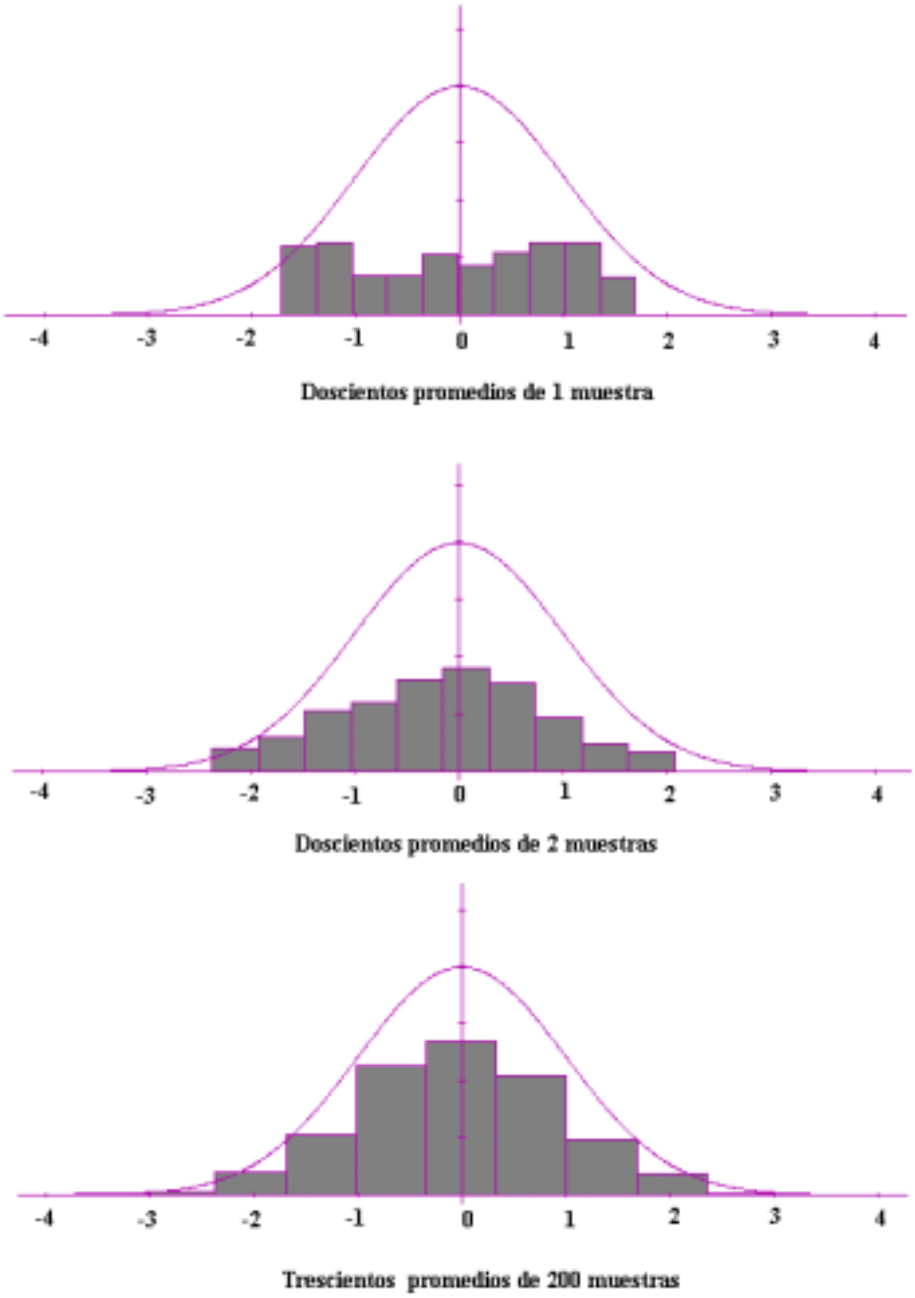


Fig. 3: Histogramas seriados de la variable aleatoria "promedios normalizados de n licencias de taxis" para n=1, 2 y 200.

Se ve luego en la Fig. 3, la manera ostensible con la que se acampanan los histogramas de la variable muestral normalizada, a medida que aumenta el tamaño muestral.

## 6. Un Intervalo de Confianza Aproximado para la Población Porteña de Taxis *N*.

Supongamos que por algún método, no necesariamente a través del promedio, se logre estimar la población de taxis mediante una estimación aproximada $\overline{N}$ . Al mismo

tiempo, aplicando rigurosamente el cálculo de probabilidades, demostramos que aceptando un margen de error $\Delta$, el verdadero valor de la cantidad de taxis *N* está con probabilidad *p* en el intervalo cerrado $I = [\ \overline{N} - \Delta, \overline{N} + \Delta\ ]$. Entonces, la terna $\{ I, \Delta, p \}$ se llama un intervalo de confianza para *N*.

Exijamos ahora que en nuestra estimación para *N* no queremos equivocarnos por más de 5000 coches, con probabilidad de alrededor del 95%. Esto querrá decir que tendremos que tener una idea de cuántas licencias *n* habrá que recolectar para que con *probabilidad del orden de 95% se verifique:*

$$\left| \mu_n - \frac{N+1}{2} \right| < 2500 \qquad (17)$$

Pero esta desigualdad (17) es equivalente a estimar para cuántos *n* esta otra desigualdad se verifica con una probabilidad de alrededor de 95% :

$$\left| \mu - \frac{N+1}{2} \right| \cdot \frac{\sqrt{n}}{\sqrt{Var(x)}} < 2500 \frac{\sqrt{n}}{\sqrt{Var(X)}} \qquad (18)$$

Finalmente, para valores grandes de *n*, y en virtud del Teorema Central del Límite, el primer miembro de la desigualdad (18) es una variable aleatoria normalizada cuya distribución es aproximadamente normal. Entonces, la indagación en la que estamos involucrados, se resuelve estimando cuál será la cantidad *n* de licencias a muestrear de modo tal que con probabilidad cercana al 95% se verifique

$$|Z| < 2500 \frac{\sqrt{n}}{\sqrt{Var(L)}} \qquad (19)$$

donde ahora *Z* es la clásica variable aleatoria normal de valor medio 0 y varianza 1. Como se sabe, las probabilidades asociadas con esta variable están extensamente tabuladas. De hecho, debe tratarse de la tabla de probabilidad más difundida y conocida de todas las existentes.

Quedan algunas dudas razonables. Primero ¿Qué se entiende por un valor grande de *n*? Respuesta: los textos [2], [6], señalan reiteradamente que una muestra de 30 elementos ya es apta. Segundo: ¿Cuánto vale la varianza $Var(L)$? Respuesta: no se sabe, y entonces hay que estimarla. Afortunadamente, la raíz cuadrada de la varianza muestral -o dispersión muestral-, tiene la agradable propiedad de que tiende a estabilizarse a medida que *n* crece, tal como lo hemos observado en la Fig. 1.

Ahora bien: nuevamente en la esquina de Bilinghurst y Corrientes, se tabularon unas 128 licencias, que arrojaron las siguientes estimaciones:

$$u_{128} = 20368.29$$

$$\sqrt{Var(L_{128})} = 10454.97$$

Buscando en una tabla normal , se ve que una variable aleatoria *N(0,1)* está entre – 1.96 y +1.96, con una probabilidad mayor al 95%:


Fig 4: Una variable N(0,1) existe con probabilidad mayor a .95 (área gris) entre -1.96 y +1.96.

Luego, si miramos (20) , podemos calcular la cantidad mínima *n* de licencias necesarias para estimar (con probabilidad mayor a .95) el promedio poblacional cometiendo un error menor que 2500 coches, y por lo tanto *la población total N con error menor que 5000 unidades.* Para tal fin despejamos *n* de

$$2500\ \frac{\sqrt{n}}{10454.97} = 1.96$$

de donde se obtienen un valor de *n* cercano a las 70 licencias. Estamos cubiertos, pues por precaución relevamos previamente 128 coches.

Por lo tanto, suponiendo que todas las licencias estén habilitadas, que no haya licencias mellizas ni *truchas*, puede afirmarse que la cantidad de taxis de la Ciudad de Buenos Aires está aproximadamente, con probabilidad superior a .95, entre

$N_{\text{inf}} = (40736\text{-}5000)$ = **38.236 coches**

y

$N_{\text{sup}} = (40736 + 5000) =$ **43.236 coches**

**7. Conclusión y Breves Comentarios Históricos.**

Con un poco de imaginación, y trabajando en equipo durante algunas semanas, un pequeño conjunto de personas entusiastas *puede estimar exitosamente la cantidad de dinero circulante que existe en el país*. Esto se conseguirá construyendo además el correspondiente intervalo de confianza, sin tener la necesidad de creer en la palabra de los funcionarios públicos del ramo. La tarea requiere anotar con cuidado los números de serie de cada billete, a razón de unas 30 anotaciones por cada denominación, y adecuar luego convenientemente los cálculos aquí realizados.

Tal como lo expresa la literatura existente, el Teorema Central del Límite ha generado un fecundo tema de trabajo y reflexión para muchos matemáticos y filósofos de la ciencia durante todo el transcurso del siglo XX. Una versión amplia de este resultado, fue presentada por Lindeberg hacia 1922, e independientemente

un poco más tarde, por Lévy en 1925. De hecho, la literatura se refiere frecuentemente al TCL como "Teorema de Lindeberg-Lévy" [2]. Sin embargo, bajo hipótesis más restrictivas, el teorema ya era conocido previamente por Liapunov y otros autores [4]. Con el correr del tiempo, fue generalizado y extendido a diferentes contextos teóricos. En nuestro país, hacia 1950, se produjeron una serie de trabajos originales de alto nivel vinculados con el mismo tema. Fueron desarrollados en el ámbito del viejo edificio de la Facultad de Ciencias Exactas, sobre la calle Perú, por el Ing. Roque Scarfiello y por el recordado y querido maestro don Alberto González Domínguez [5].

**Referencias.**


1) Cochran, W. G.: *Sampling Techniques*, New York, Wiley & Sons, 1978, pgs. 18-45. Se ofrecen y se justifican rigurosamente las técnicas profesionales de muestreo y estimación.
2) Dudewicz, E., Mishra, S. *Modern Mathematical Statistics*, ídem anterior, 1998. Se demuestra una versión básica del TCL y se indican algunas generalizaciones.
3) *Enciclopaedia Britannica*, London, CD Edition, 1999, artículos: *World War II*, *Radar*, y *Microwave*.
4) Feller, W: *An Introduction to Probability Theory and its Applications*, John Wiley & Sons, 1968, pgs. 212-233. Esta obra contiene los teoremas básicos sobre medias y varianzas de variables aleatorias discretas, tales como los utilizados aquí.
5) González Domínguez, A., Scarfiello, R.:*"Teoremas Límites para Productos de Variables Aleatorias"*, Buenos Aires, Contribuciones Científicas, FCEFyN, UBA, 1950.
6) Rohatgi, V.: *Statistical Inference*, New York, John Wiley & Sons, 1984: El problema de los taxis está explícitamente tratado en varios capítulos, en donde se construyen intervalos de confianza basados en diferentes estimadores.


Ricardo Miró, para http://www.rinconmatematico.com
Si quieres transmitirnos alguna inquietud generada por este u otro artículo, puedes hacerlo en http://www.rinconmatematico.com/foros