El Análisis Matemático aplicado al CÁLCULO DE LA MUESTRA

El tamaño de la muestra es (in)finito

Se realiza un estudio sobre la función que describe el tamaño de la muestra cuando el proceso es aleatorio simple, y se analiza su comportamiento cuando aumenta la población. Tal estudio se fundamenta en la aplicación de las herramientas del análisis matemático, específicamente la teoría de límites para una función real, y el cálculo diferencial en funciones de una variable. Si bien es cierto, existe la idea generalizada de que a medida que aumenta la población, el tamaño de la muestra aumenta en la misma proporción; pero en este artículo, se llega a demostrar que contrariamente a lo que se piensa, a pesar de que la población aumente, el ritmo de crecimiento de la muestra, disminuye; haciendo que el tamaño de la muestra converja a un valor finito. Por tal razón, en los procesos investigativos, donde se tiende a emplear únicamente este tipo de muestreo, se debe hacer hincapié que su aplicación tiene determinadas condiciones que la restringen.

POR: Dr. Miguel Angel Reinoso Sánchez

Universidad Estatal de Milagro,
Departamento de Investigación

E-mail
mreinosos@unemi.edu.ec

Introducción

El presente artículo surge de una conversación en torno al proceso de selección de la muestra para un estudio socioeconómico que lleva a cabo el Departamento de Investigación. En todo proceso investigativo que demanda tomar una muestra para su análisis, aparecen varias interrogantes, entre ellas: ¿en qué medida se puede asegurar que la muestra escogida es representativa de la población?; es decir, cómo garantizamos que el estudio realizado sobre la muestra constituya la realidad latente de la población. Y una segunda pregunta, tiene que ver con la cantidad de elementos a ser considerados, se asume que si la población aumenta, también aumenta el tamaño de la muestra; en ese sentido, ¿si la población crece indefinidamente, entonces la muestra debería ser también infinita?

La primera interrogante tiene su respuesta en la estadística inferencial, mientras que la segunda puede ser resuelta utilizando las herramientas del análisis matemático, que si bien es cierto, constituyen elementos básicos para la formación de un ingeniero o investigador, su poder radica en que mediante su empleo, podemos interpretar, extrapolar, conjeturar y hasta estudiar las limitaciones de un determinado modelo[1].

Esta es la razón misma del artículo, demostrar a lectores, docentes y fundamentalmente a estudiantes, que con los conocimientos adquiridos en las aulas, así sean estos muy elementales, es posible dar aportes al conocimiento científico, construir pequeños peldaños que hacen que nuestro propio "edificio matemático" gane en altura.

Con estos antecedentes, realizaremos un análisis de la fórmula que comúnmente se emplea en proyectos de investigación para determinar el tamaño de la muestra, pero en muy pocos se menciona que su aplicación tiene algunas restricciones, como por ejemplo: el tipo de muestreo es aleatorio simple, se conoce con certeza el tamaño de la población, se la utiliza para sacar una proporción de la población más no para hacer un análisis de medias.

Metodología

De acuerdo a la bibliografía [1], [2], en un muestreo aleatorio simple, el tamaño de la muestra está definido dependiendo de la población, si ésta es finita o infinita:

[1] Se entiende por modelo, la descripción matemática de una situación real, en ese sentido, existen modelos matemáticos empleados en todo ámbito del desarrollo humano, desde las ciencias sociales hasta las ciencias de la vida.

[1] Doctor en Física, Escuela Superior Politécnica del Chimborazo. Ex director de la comisión de Investigación y Experimentación de la Unidad Educativa Nº 3, de la Fuerza Aérea Ecuatoriana, Taura. Ha elaborado trabajos en el campo de la enseñanza de ciencias exactas, investigación aeroespacial, física nuclear y estudios sobre los efectos de radiaciones ionizantes en seres vivos.

Cuando la población es finita y se conoce con certeza su tamaño:

$$n = \frac{N\,p\,q}{\frac{(N-1)\,E^2}{Z^2} + p\,q} \qquad (1)$$

Cuando la población es infinita:

$$n = \frac{Z^2\,p\,q}{E^2} \qquad (2)$$

donde:

n: tamaño de la muestra,
N: tamaño de la población,
Z: nivel de confianza,
p: posibilidad de ocurrencia de un evento,
q: posibilidad de no ocurrencia de un evento, q = 1 – p,
E: error de la estimación.

Para el tratamiento de este modelo matemático, se va a realizar lo siguiente [4], [5]:

a. Graficar la relación entre el tamaño de la muestra en función del tamaño de la población, considerando a los parámetros Z, p y E fijos

b. Analizar la monotonía de la función n = f(N); es decir basados en las herramientas del análisis matemático encontrar los intervalos donde n es creciente y decreciente; y adicionalmente los puntos críticos .

c. Analizar la variación de la razón entre el tamaño de la muestra y el tamaño de la población, esta proporción se la puede expresar mediante porcentajes y a partir de sus resultados emitir algunas conclusiones.

d. Establecer la misma variación anterior pero para diferentes valores de Z, p y E.

e. Demostrar que la fórmula (2) es un resultado particular de la fórmula (1), como es de suponerse si la población es infinita, al evaluar en la ecuación (1) el límite de la función f(N) cuando N tiende al infinito, por fuerza debe darnos la ecuación (2).

Desarrollo

Si bien es cierto, el tamaño de la muestra (que en adelante llamaremos simplemente "n") depende de varias variables, vamos a especificar la dependencia de una sola, ésta es el tamaño de la población "N"; y consideramos a p, E y Z como parámetros.

Habitualmente se trabaja con un error de la estimación de 0,05 (E=0,05), y un nivel de confianza correspondiente al 95%, que sería de 1,96; es decir Z=1,96.

Los valores de p se encuentran comprendidos en un intervalo entre cero y uno, pero la fórmula es simétrica alrededor del 0,5 (reflejada en la expresión pq, donde q=1-p); entonces un valor de p=0,3; tendrá un resultado equivalente al p=0,8. Por ende, basta analizar la función para valores de p entre 0 y 0,5

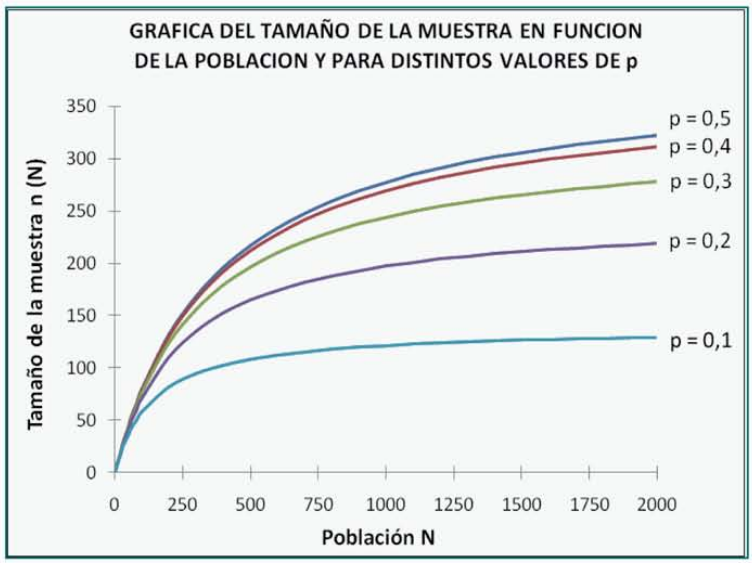

Figura 1: Gráfica del tamaño de la muestra en función de la población.

Con este antecedente, en la Figura 1. se presentan las gráficas de la función correspondiente al tamaño de la muestra n=f(N) para diversos valores de p; al observar las gráficas uno podría pensar que "siempre" esta función es creciente, pero esta apreciación no es del todo cierta; *las gráficas no son el instrumento para obtener resultados concluyentes o demostrar un teorema*, simplemente sirven como una ayuda para orientar el proceso de análisis o demostración.

Formalmente hablando, si se quiere determinar en qué intervalos la función es creciente y en qué intervalos es decreciente (monotonía de la función), se debe aplicar el criterio de la primera derivada, que dice:

"Si n` > 0 entonces la función n es estrictamente creciente; caso contrario, si n` < 0 entonces la función n es estrictamente decreciente"'1

Donde n` representa a la primera-derivada de n con respecto a N, es decir:

$$n' = \frac{d\,n}{d\,N} \qquad (3)$$

Por facilidad realizamos la siguiente asignación:

$$\left(\frac{E}{Z}\right)^2 = a \qquad (4)$$

Quedando la ecuación (1) como:

$$n = \frac{N\,p\,q}{a\,(N-1) + p\,q} \qquad (5)$$

Es así que la primera deriva es:

$$n' = \frac{p\,q\,(pq - a)}{\left[a\,(N-1) + p\,q\right]^2} \qquad (6)$$

Para determinar el o los intervalos donde la función es creciente, planteamos la inecuación:

$$n' = \frac{p\,q\,(pq - a)}{\left[a\,(N-1) + p\,q\right]^2} > 0 \qquad (7)$$

De lo que podemos apreciar, tanto pq y el denominador serán positivos, por ende, para que la desigualdad (7) sea válida, debe cumplirse que:

$$p\,q - a > 0$$

es decir,

$$p\,q > a \qquad (8)$$

Aunque el valor de a es muy pequeño, casi nulo, esta desigualdad no se satisface en todo el intervalo [0, 1]; lo que quiere decir que la función efectivamente es CRECIENTE para casi todos los posibles valores de p, pero hay una región en que la función n es decreciente.

Recordando que q = 1 – p, la expresión (8) se transforma en:

$$-p^2 + p - a > 0 \qquad (9)$$

Planteándose de esta manera una inecuación de segundo grado, que puede resolverse mediante el estudio del signo: determinando las raíces del polinomio, graficando el polinomio en términos del signo de p^2 y estableciendo el intervalo que satisface la condición (9), misma que indica que el polinomio debe ser mayor que cero, es decir: positivo (+).

Las raíces del polinomio son:

$$p_{1,2} = \frac{-1 \pm \sqrt{1-4a}}{-2} \qquad (10)$$

Reemplazando los valores habituales de E=0,05 y Z=1,96; en la fórmula (4) tenemos:

$$a = 0{,}000\,651$$

Con este valor obtenemos las raíces de p:

$$p_1 = \frac{-1 + 0{,}998697}{-2} = 0{,}000651\ldots$$

$$p_2 = \frac{-1 - 0{,}998697}{-2} = 0{,}999348\ldots$$

Como el signo de p^2 en la inecuación (9) es negativo, la concavidad de la parábola es hacia abajo, quedando como resultado que la inecuación se satisface cuando el signo del polinomio es positivo (Figura 2).

Por lo tanto, la función n=f(N) es estrictamente creciente cuando p toma valores entre 0,000651 y 0,999348, y por fuera de este intervalo, la función es decreciente; es decir:

Si $p \in (0{,}000651\ldots;\ 0{,}999348\ldots)$ entonces la función n=f(N) es estrictamente creciente[3].

Pues bien, si consideramos un valor de p menor que 0,000651; por ejemplo: p=0,00065; efectivamente la función es decreciente, pero sin importar el tamaño de la población, la muestra contaría a lo sumo con un elemento (ver figura 3)!!!

Exceptuando este tipo de casos (que abarca menos del 0,06% de los posibles valores de p), en el resto del intervalo, la función del tamaño de la muestra es CRECIENTE, lo cual significa que a medida que aumenta la población, aumentará el tamaño de la muestra.

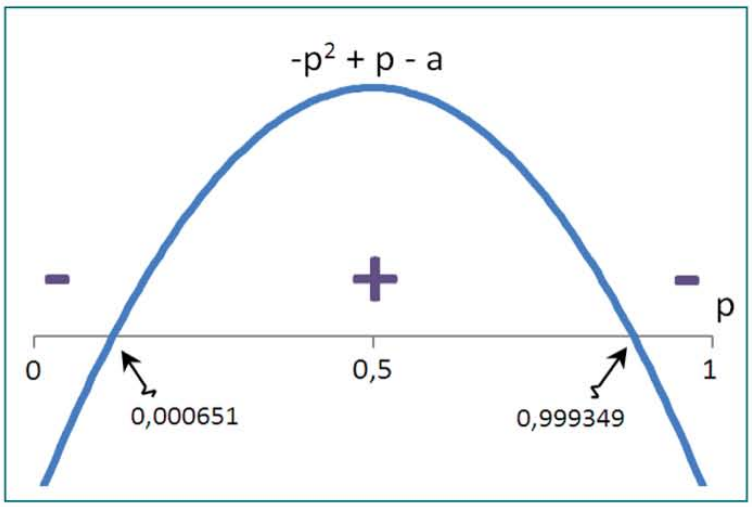

Figura 2: Estudio del signo del polinomio - p^2 + p - a

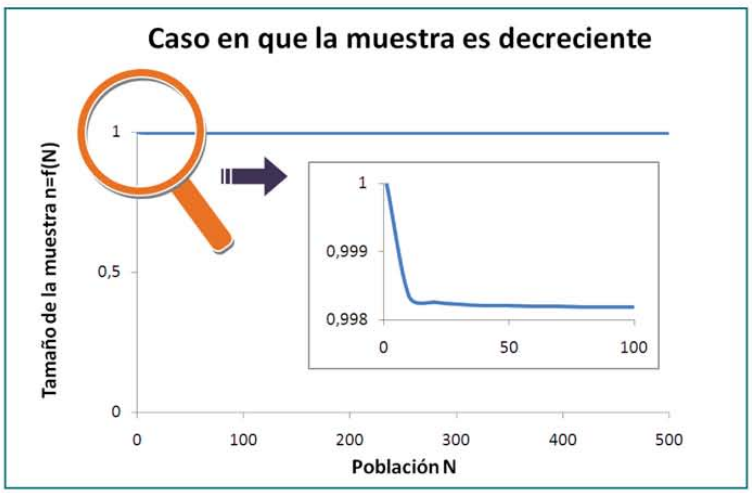

Figura 3: Para p=0,00065 la función n=f(N) es decreciente, pero el tamaño de la muestra sería de apenas un elemento.

Pero, lo interesante está en que el aumento de n no va en la misma proporción del aumento de la población; más bien, resulta ser todo lo contrario. Por ejemplo, si la población es de 100 elementos, la muestra contiene a 80 (la proporción es del 80%); y si la población es de 1 000, la muestra contiene a 278 elementos (el tamaño de la muestra aumentó, pero la proporción se reduce al 27%)

Este hecho sugiere que debemos analizar la razón entre el tamaño de la muestra y el de la población, para ello planteamos la función "r":

$$r(N) = \frac{n}{N} \qquad (11)$$

Empleando la fórmula (5) del tamaño de la muestra, queda que:

$$r(N) = \frac{p\,q}{a\,(N-1) + p\,q} \qquad (12)$$

Al sacar la primera derivada:

$$r' = -\frac{a\,p\,q}{\left[a\,(N-1) + p\,q\right]^2}$$

Y sabiendo que a, p, q y el denominador (por estar elevado al cuadrado) serán positivos, el signo menos hace que la primera derivada sea siempre negativa; luego, "la función r(N) resulta ser decreciente en todo el intervalo asociado a p".

Quiere decir que, independientemente del valor de p a medida que aumenta la población, la proporción que se toma para su estudio disminuye[4].

[3] Recordar que esta conclusión es válida para E=0,05 y Z=1,96.

[4] Recalcamos que se trata de un muestreo aleatorio simple.

Graficando la función r(n) para distintos valores de p, obtenemos (ver figura 4).

Este hecho de que el crecimiento de la muestra no va al mismo ritmo que el crecimiento de la población, o que la relación es decreciente con tendencia de llegar a cero, nos hace pensar que la función n=f(N) tiene una asíntota horizontal; es decir existe una cota (extremo) superior que el tamaño de la muestra no puede sobrepasar; pero claro está, puede ocurrir que tal asíntota no exista, y que el crecimiento de la muestra sea infinito (ver figura 5).

Para saber si existe o no la asíntota, analicemos cuál será el límite de la función n cuando el tamaño de la población aumente al infinito [6]:

$$\lim_{N\to\infty} n = \lim_{N\to\infty} \frac{N\,p\,q}{\frac{(N-1)E^2}{Z^2} + p\,q}$$

Al ser un límite al infinito, dividimos tanto el numerador como denominador para N:

$$\lim_{N\to\infty} \frac{p\,q}{\frac{E^2}{Z^2} - \frac{E^2}{N\,Z^2} + \frac{p\,q}{N}} = \frac{p\,q}{\frac{E^2}{Z^2}}$$

La última igualdad se obtuvo evaluando el límite; pues como $N\to\infty$, tanto la expresión $\frac{E^2}{N\,Z^2}$ como $\frac{p\,q}{N}$ tienden a cero; quedando únicamente en el denominador $\frac{E^2}{Z^2}$

Por consiguiente, existe una cota superior para la función n(N), y su cota está dada por:

$$\lim_{N\to\infty} n = \frac{Z^2\,p\,q}{E^2} \qquad (13)$$

Resultado final que coincide exactamente con la fórmula (2) del tamaño de la muestra cuando la población es infinita.■

Así que la función n=f(N) efectivamente tiene una asíntota horizontal, cuyo valor dependerá del nivel de confianza Z y del error de la estimación E.

En el ejemplo que hemos analizado; para Z=1,96; E=0,05 y para distintos valores de p; se presentan los diferentes tamaños de la muestra (Tabla 1.) y el valor máximo o cota superior cuando la población es infinita (Tabla 2.):

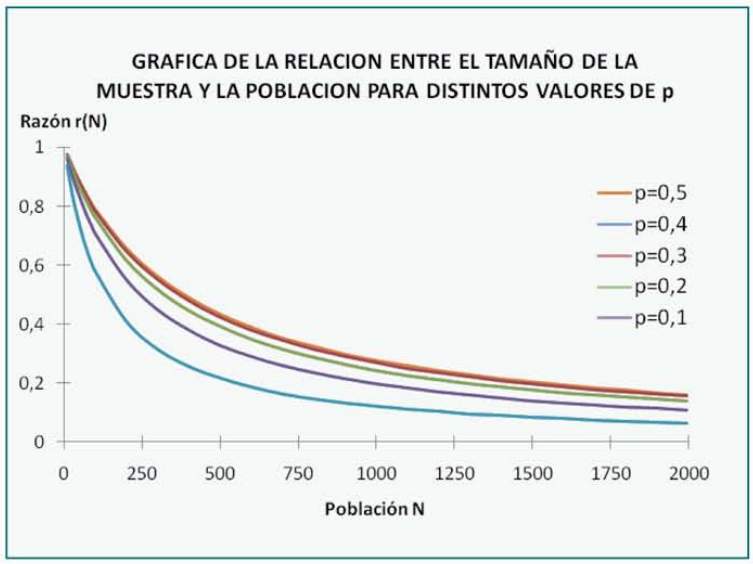

Figura 4: Gráfica de la relación entre el tamaño de la muestra con respecto a la población.

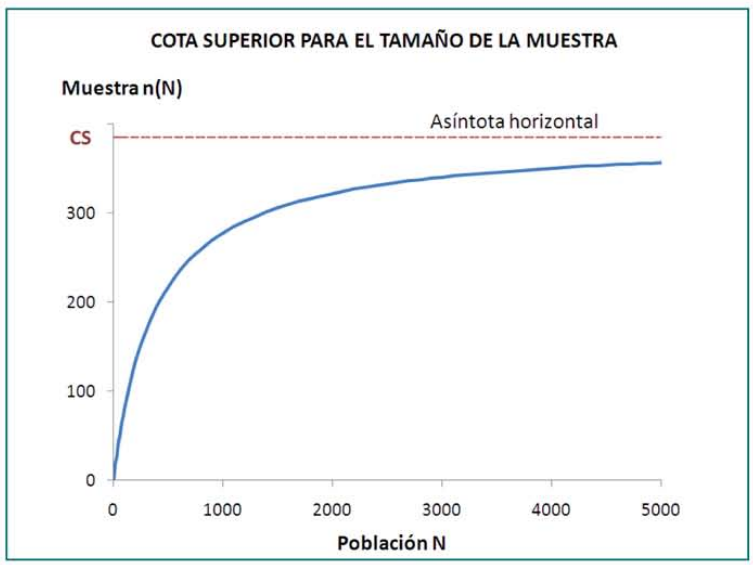

Figura 5: Cota superior para el tamaño de la muestra con p=0,5.

Valor de p	Tamaño máximo de la muestra
0,1	138
0,2	246
0,3	323
0,4	369
0,5	384
0,6	369
0,7	323
0,8	246
0,9	138

Tabla 2: Valores máximos de la muestra, para Z=1,96 y E=0,05
Fuente: Propia

Conclusiones

El análisis realizado en el presente artículo se basa en la variación de las proporciones p y q; pero también se lo puede enfocar al caso de muestreos pilotos con relación al cambio de varianzas o desviaciones estándar muestrales y poblacionales.

Hemos demostrado que la fórmula (2) correspondiente al tamaño de la muestra cuando la población es infinita, es consecuencia de la fórmula (1) que describe el tamaño de la muestra para una población finita, simplemente evaluando el límite cuando $N\to\infty$.

El hecho de realizar tal demostración por otra vía distinta a la Estadística Inferencial, consolida aún más la estructura matemática en la que se

Tabla 1: Tamaño de la muestra para distintos valores de p; con Z=1,96 y E=0,05*

Población N	Valores de p								
	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
0	0	0	0	0	0	0	0	0	0
10	9	10	10	10	10	10	10	10	9
20	18	19	19	19	19	19	19	19	18
30	25	27	28	28	28	28	28	27	25
40	31	35	36	36	36	36	36	35	31
50	37	42	43	44	44	44	43	42	37
60	42	48	51	52	52	52	51	48	42
70	47	55	58	59	59	59	58	55	47
80	51	61	64	66	66	66	64	61	51
90	55	66	71	73	73	73	71	66	55
100	58	71	77	79	80	79	77	71	58
200	82	111	124	130	132	130	124	111	82
300	95	135	156	166	169	166	156	135	95
400	103	153	179	192	196	192	179	153	103
500	109	165	196	212	217	212	196	165	109
600	113	175	210	229	234	229	210	175	113
700	116	182	221	242	248	242	221	182	116
800	118	188	230	253	260	253	230	188	118
900	120	193	238	262	269	262	238	193	120
1000	122	198	244	270	278	270	244	198	122
2000	129	219	278	312	322	312	278	219	129
3000	132	227	291	329	341	329	291	227	132
4000	134	232	299	338	351	338	299	232	134
5000	135	234	303	344	357	344	303	234	135
6000	135	236	306	347	361	347	306	236	135
7000	136	238	309	350	364	350	309	238	136
8000	136	239	310	353	367	353	310	239	136
9000	136	239	312	354	368	354	312	239	136
10000	136	240	313	356	370	356	313	240	136

* Nótese que cuando aumenta la población, el tamaño de la muestra tiende a estabilizarse y que para poblaciones muy grandes la fórmula genera una muestra que ya no es representativa.
Fuente: Propia

fundamenta, pues siempre es importante que los resultados de una teoría sean sustentados por otras.

Como el tamaño de la muestra está acotado, a medida que se consideren poblaciones más grandes, la proporción que se escoge de la población para su estudio, va reduciéndose; significa que llegará un punto en el que la muestra que tomemos no sea significativa, lo cual conduce a decir que el análisis desarrollado sobre la muestra no corresponde a la realidad de la población.

Por esta razón existen otro tipo de muestreos, como por ejemplo: muestreo estratificado, sistemático, por estadíos múltiples, por conglomerados, etc.; cada uno de ellos se aplica bajo determinados criterios, de ahí que es importante establecer las condiciones iniciales de un problema a ser analizado para poder escoger adecuadamente la técnica que debamos emplear en su estudio.

Finalmente, se puede establecer en qué medida son válidas las recomendaciones emitidas por investigadores, cuando mencionan que una muestra representativa de la población debe tener el 30% de sus elementos.

Si deseamos confirmar la validez de este argumento, podríamos calcular el valor medio de la fórmula r (que mide proporciones) y verificar si nos da el 30%; caso contrario se puede evaluar hasta qué punto es recomendable usar dicha recomendación.

Referencias bibliográficas

[1] Anderson, D., Sweeney, D. y Williams, T. (2007). Estadística para Administración y Economía, Thomson, México, pp. 303-308.
[2] Devore, J. (1998). Probabilidad y Estadística para ingeniería y ciencias, Thomson, México, p. 276.
[3] Lagares, P. y Puerto, J. (2001). Población y muestra. Técnicas de muestreo. Extraído el 25 de junio de 2009, del sitio http://optimierung.mathematik.unikl.de/mamaeusch/veroeffentlichungen/ver_texte/sampling_es.pdf
[4] Lara, J. (2001). Análisis matemático, Universidad Central del Ecuador, Quito, p. 120.
[5] Swokowsky, E. (1989). Cálculo con geometría analítica. Grupo editorial Iberoamericana, México, pp. 175-178
[6] Sydsaeter, K. y Hammond, P. (1196). Matemáticas para el análisis económico. Prentice Hall, Madrid, pp. 142, 143.