Base de datos

TP2

Departamento de Computación
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

NN_3

Integrante	LU	Correo electrónico
Sergio González	723/10	sergiogonza90@gmail.com
Gino Scarpino	392/08	gino.scarpino@gmail.com

Reservado para la cátedra

Instancia	Docente	Nota
Primera entrega		
Segunda entrega		

Índice

1. Introducción 3

2. Estimadores 3

3. Análisis de métodos 4

3.1. Distribuciones utilizadas 4

3.1.1. Distribucion normal (Ejemplos) 4

3.1.2. Distribucion uniforme (Ejemplos) 6

3.2. Análisis de los estimadores: Parámetro fijo 8

3.3. Análisis de los estimadores: Parámetro variable 10

3.3.1. Distribución uniforme 10

3.3.2. Distribución normal 11

1. Introducción

Es fundamental para cualquier motor de bases de datos, poseer un planificador para resolver consultas lo más eficiente posible. Medir el costo de un método de evaluación puede ser muy complejo, pero como se indica en el paper de Piatetsky-Shapiro, aproximadamente son las cantidad de operaciones de entrada y salida en disco que el motor realiza. Por eso mismo, se trata de minimizar estas operaciones.

Una de las formas de poder minimizar estas operaciones, es poder conocer aproximadamente cual puede ser la distribución de un set de datos, y de esta forma poder estimar cuantas tuplas se obtendrá por el hecho de aplicar un filtro (WHERE) u otro. El hecho de poder minimizar las tuplas resultantes que se obtendrán en una consulta, puede hacer que al momento de materializar la misma (por ejemplo en caso de tener que hacer un JOIN posterior) haga que las bajadas a disco de las tuplas se minimizen considerablemente.

Si bien computar la distribución exacta de un set de datos puede ser muy costoso, existen métodos con el cual se puede aproximar dichas distribuciones y de esa forma poder decidir cual es el mejor camino a seguir al momento de tener que resolver una consulta.

2. Estimadores

En este trabajo veremos tres estimadores, que los explicaremos a continuación:

- **Histograma Clásico:** este estimador divide el rango de los valores en varios subrangos llamados *buckets*. Contabiliza los cada valor aumentando la cantidad del *bucket* correspondiente. Se basa fuertemente en estimar la probabilidad de un valor v con la cantidad total del *bucket* correspondiente a ese v. Cuanto mayores subrangos haya, máyor va a ser la precisión para determinar la frecuencia del valor y de ahí estimar su selectividad.

- **Pasos Distribuidos:** estimador inventado por Piatetsky-Shapiro. Al igual que en los histogramas clásicos, usa el concepto de *bucket* pero en vez de usar un ancho (rango) fijo para cada bucket, la cantidad de elementos en un *bucket* está determinada por la altura del mismo. Depende de la cantidad de *buckets* que se desea, dividiendo la cantidad total de elementos (tuplas) por esa cantidad se obtiene la altura.

- **Estimador Propio:** Nos basamos en el histograma clásico pero introducimos una pequeña mejora (más adelante se verá en los análisis de los testeos). Realizamos un histograma clásico, luego detectamos una cantidad arbitraria de buckets con mayor cantidad de elementos y dividimos su rango en la mitad creando dos nuevos buckets. Cabe destacar que esto no significa que al dividir el rango en la mitad, se dividan la cantidad de elementos en la mitad. Para poder realizar esto, utilizamos un arreglo que mantiene información específica de cada *bucket*, su rango. Notamos que la desventaja con respectoa los otros dos estimadores es que en cada subdivisión de un buckets necesitamos realizar una consulta a la base de datos para determinar cuántos elementos van en cada nuevo bucket. Para este trabajo, creamos el histograma clásico con la mitad de buckets pasados como parámetro. Vamos dividiendo buckets que tengan el máximo número de elementos hasta llegar a la cantidad de buckets indicado por parámetro.

3. Análisis de métodos

3.1. Distribuciones utilizadas

3.1.1. Distribucion normal (Ejemplos)

La distribución normal es una de las distribuciones que mas aparece en la vida real. A continuación se presentan 2 ejemplos de la misma.

Altura de una persona

Es ampliamente conocido que los caracteres morfológicos de individuos, tales como la estatura, siguen el modelo normal en todo el mundo. A simple vista, se puede pensar como por lo general la altura de las personas suele estar entre 1-70 y 1.75 metros, o por lo menos en la mayoría de los casos es así. No suele haber muchas personas que midan menos de 1.50, y a la ves no hay tampoco demasiadas personas que superen los 2 metros. Haciendo un muestreo poblacional y realizando un histograma del mismo se puede visualizar esta intuición.

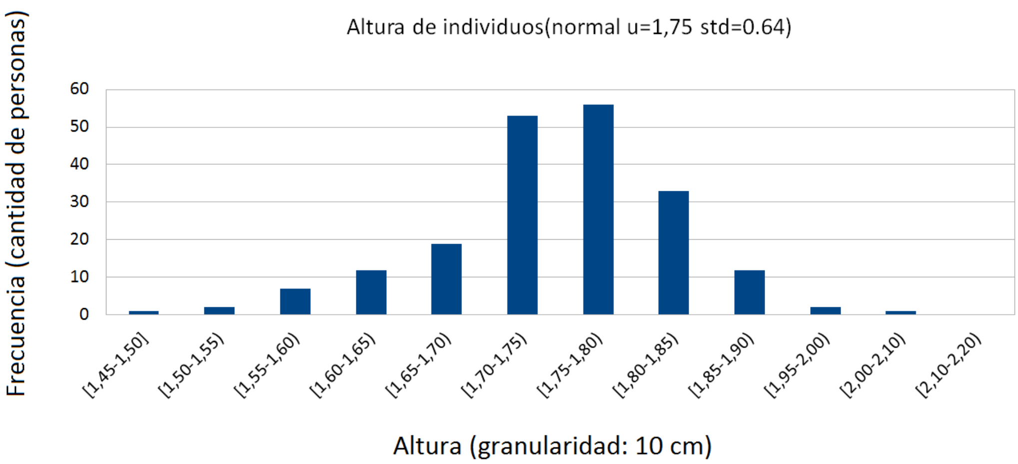

Figura 1: Histograma altura

Se puede ver como para el muestreo realizado, la mayorá de las personas caen en la altura entre 1.70 y 1.80 metros, dando como resultado aproximadamente, una normal con media 1,75 y desvío standard 0.64.

IQ de una persona

Otro ejemplo muy conocido es el de el coeficiente intelectual de las personas, también llamado IQ. Según el siguiente ranking, vemos que una inteligencia normal debería esta entre los 90 y los 109 de coeficiente intelectual, por lo que es de esperar que la mayor parte de la población este en este promedio.

IQ Range	Clasificación
130 o mas	Muy Superior
120-129	Superior
110-119	Arriba del promedio
90-109	Promedio
80-89	Abajo del Promedio
70-79	Limite
69 o menos	Extremadamente bajo

Veamos un histograma sobre el muestreo del IQ de los individuos de una población.

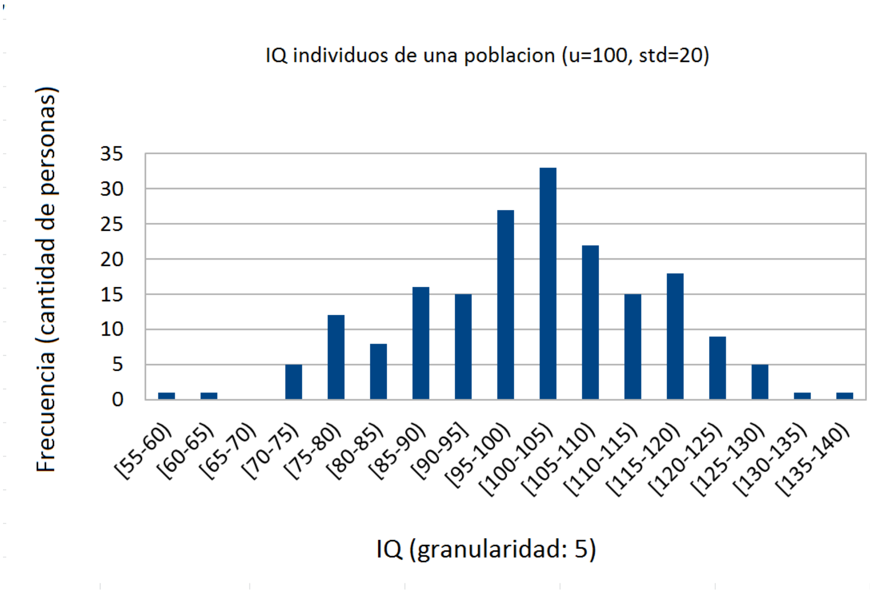

Figura 2: Histograma IQ

En la figura 2, podemos un ver como efectivamente la mayoría de la población se centra en un IQ de 100, con una desvío al rededor de 20.

3.1.2. Distribucion uniforme (Ejemplos)

La distribucion uniforme es una de las mas conocidas, y como la nomal, una de las mas presentes en la realidad. Esta distribucion es muy simple, basicamente plantea que si tenemos un espacio muestral $S = \{m_1, m_2, ..., m_n\}$, ($\forall m_i \in S$, $P(m_i) = 1/n$), osea, todos los sucesos tiene la misma probabilidad de ocurrir.

El ejemplo mas clasico para entender esa distribucion, es el lanzamiento de un dado de 6 caras (no cargado), en onde $S = \{1, 2, 3, 4, 5, 6\}$ y cada elemento tiene la misma probabilidad de salir, osea 1/6.

Tiempo de espera de un colectivo

Otro ejemplo un poco mas interesante, es si tomamos un rango de tiempo, y medimos el tiempo de espera de un colectivo en ese rango de tiempo. Si bien este ejemplo dependera mucho de sobre que linea de colectivo hagamos el muestreo, se puede tomar un rango en particular en el cual sabemos que el tiempo de espera no sera mayor o menor a eso. A continuacion se presenta un histograma sobre un dataset sobre los tiempos de cierto colectivo en el rango de 5 minutos a 30 minutos.

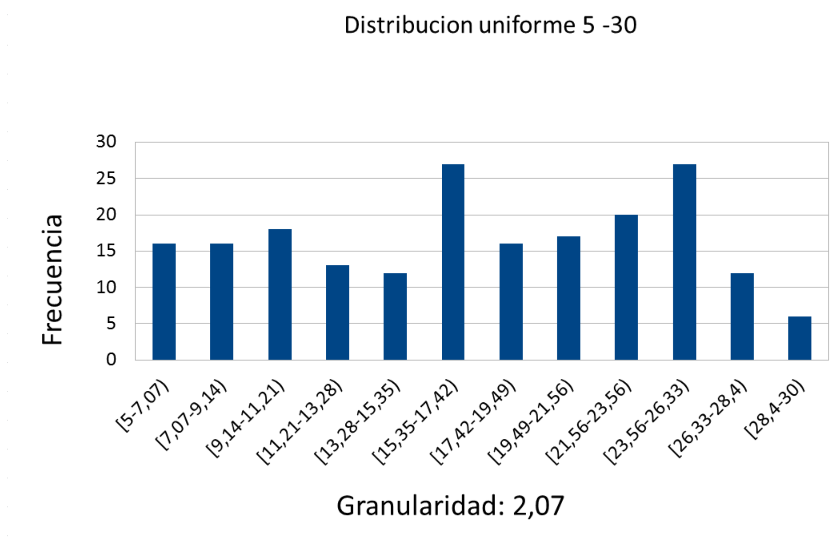

Figura 3: Histograma tiempos de espera de colectivo

3.2. Análisis de los estimadores: Parámetro fijo

Primero realizaremos un análisis sobre, un mismo set de datos, y un parámetro igual para todos los estimadores, y veremos como se comportan. De esta forma podremos analizar para las distintas distribuciones de datos, como se comporta cada método y cuales son las ventajas y desventajas de cada uno.

Intuitivamente podríamos decir que, al ser “Distribution Steps” un histograma que no tiene el mismo ancho, sino que tiene misma altura en todos los *bins*, claramente es esperable que para un parámetro fijo, y un set de datos fijo, “Distribution Steps” tenga error menor.

Para comprobar esto, utilizaremos el set de datos provisto para la materia y realizaremos gráficos comparando los 3 estimadores implementados.

Caso 1

Parametro	10
Columna	C2
Valor maximo	1002
Valor minimo	-671
Distribución	Normal, Media=200

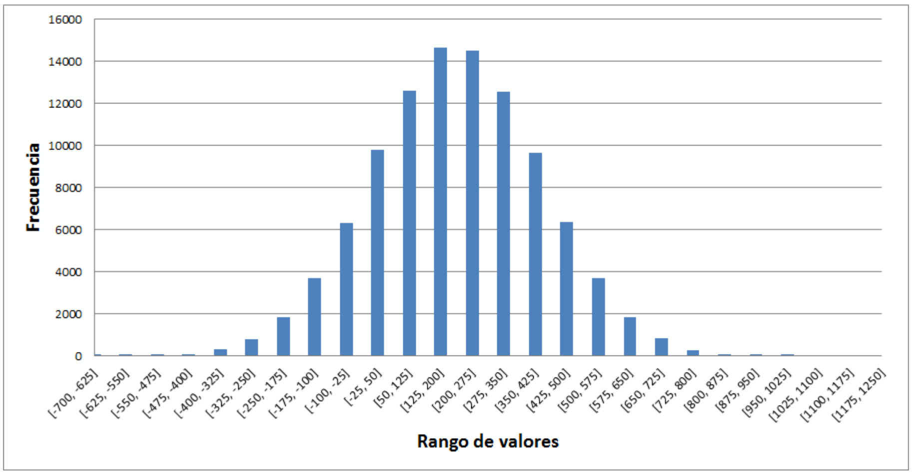

Figura 4: Grafico de distribucion de la columna C2 del set de datos de la caterda

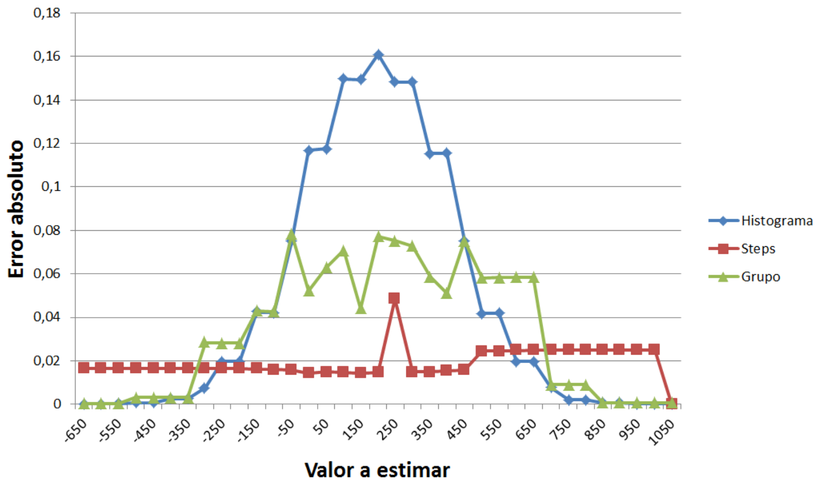

Figura 5: Errores variando valor a estimar con parametro fijo

En la figura 4 se ve la distribucion del set de datos que estamos analizando. Se puede ver como es una distribucion Normal con media 200 y un desvío alrededor de 300

En la grafico de la figura 9 se ve como, teniendo los parametros de los estimadores fijos, y variando el valor a estimar, el estimador de Distribution Steps obtiene un error constante y bastante chico en todo el rango, pero Classic Histogram se comporta mejor en los casos que estan por afuera del desvio standard de la normal (en este caso, al rededor de 300).

También se puede ver como el estimador Classic Histogram obtiene errores muy grandes cuando el valor es muy cercanos a la media. En la media misma se ve como el error llega al máximo.

En cuanto al estimador ideado por el grupo, se ve como al igual que el Histograma, obtiene errores muy chicos en valores lejos del desvio standard de la normal, y a su ves, errores altos en los valores al rededor de la media. Sin embargo, estos errores son inferiores a los obtenidos en el histograma. Esto probablemente se deba a que el estimador de Grupo, es un histograma clasico pero con una re-distribución en los *bins* con mas valores, los cuales en este caso estarÃ¡n cerca de la media.

3.3. Análisis de los estimadores: Parámetro variable

Al igual que en el análisis previo, realizamos mediciones de los estimadores variando el parámetro de entrada sobre las mismas poblaciones mencionadas.

El parámetro que se varía en todos los estimadores representa la cantidad de *buckets*, aunque en el estimador de pasos distribuidos se lo llaman *steps*.

3.3.1. Distribución uniforme

- **Operación por igualdad**

  lala

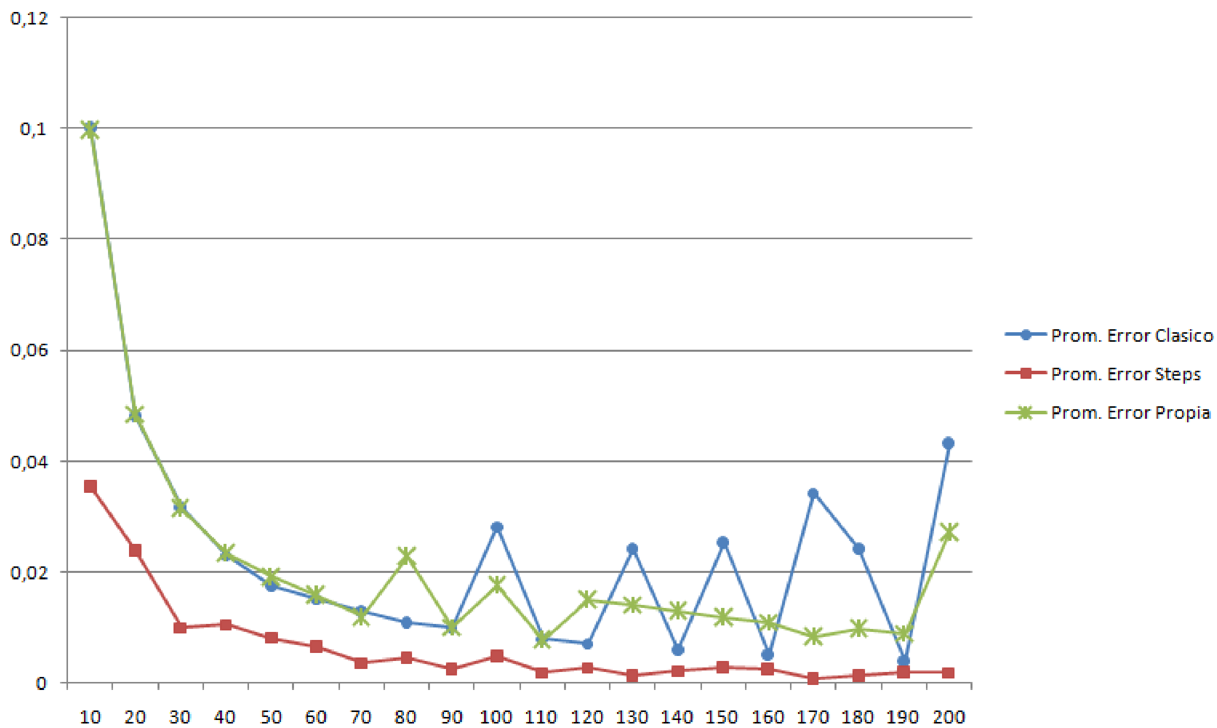

Figura 6: Error promedio de la Columna C0 de la tabla brindada por la materia

  lala

- **Operación por mayor**

  lala

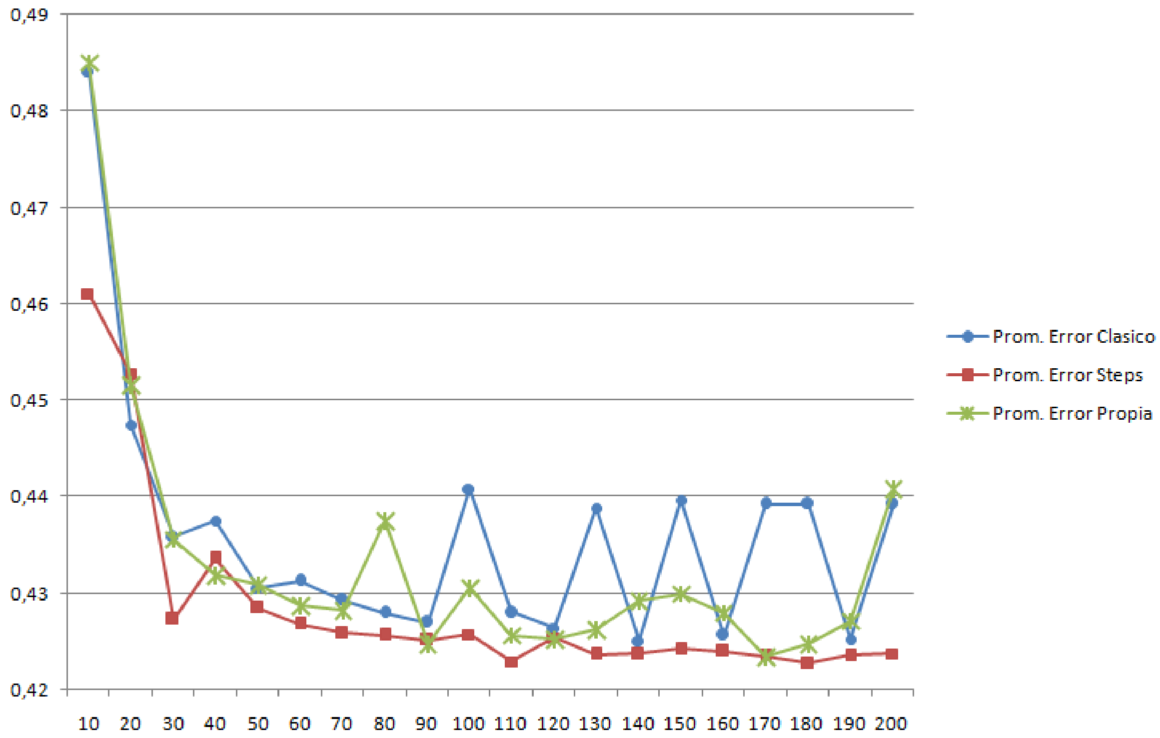

Figura 7: Error promedio de la Columna C0 de la tabla brindada por la materia

lala

3.3.2. Distribución normal

- **Operación por igualdad**

  lala

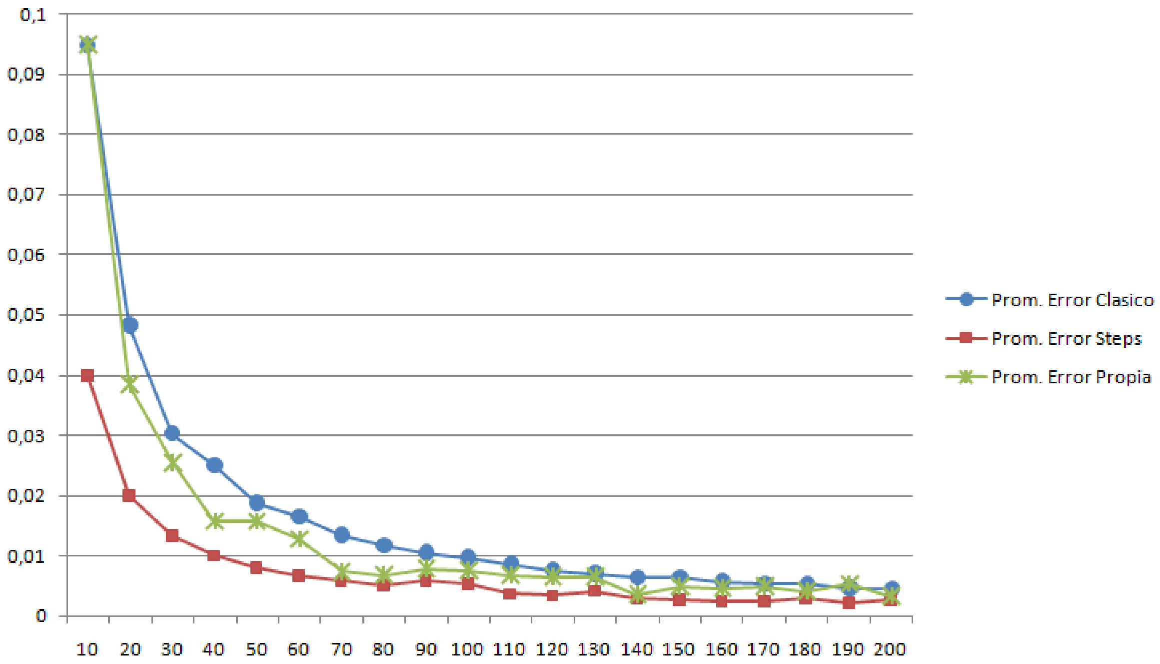

Figura 8: Error promedio de la Columna C2 de la tabla brindada por la materia

lala

- **Operación por mayor**

  lala

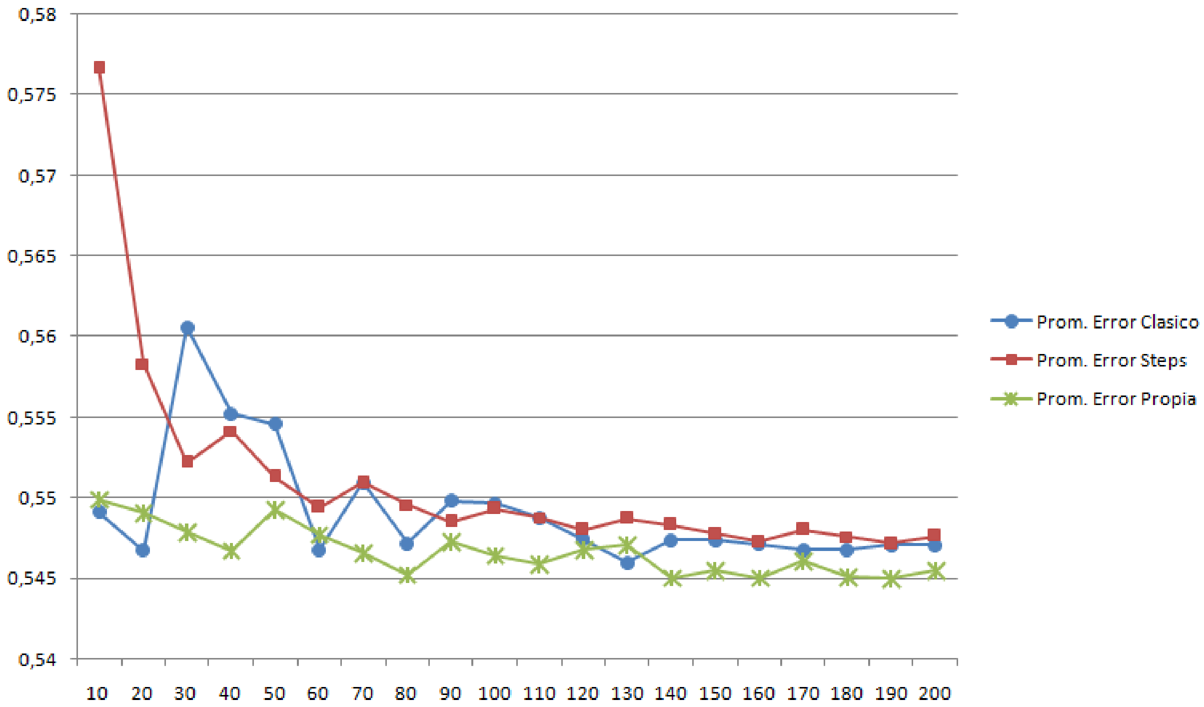

Figura 9: Error promedio de la Columna C2 de la tabla brindada por la materia

lala