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

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 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 que lo cumplan. 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 de elementos 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 para ciertos casos (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 indicada por parámetro.

3. Análisis de métodos

3.1. Distribuciones utilizadas

3.1.1. Distribución 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 las características morfológicas de individuos, tales como la estatura, siguen el modelo normal en todo el mundo. En general, se puede pensar la altura de las personas suele estar entre 1-70 y 1.75 metros, debido a que en la mayoría de los casos es así. Es minoría las personas que midan menos de 1.50, y a la vez 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.

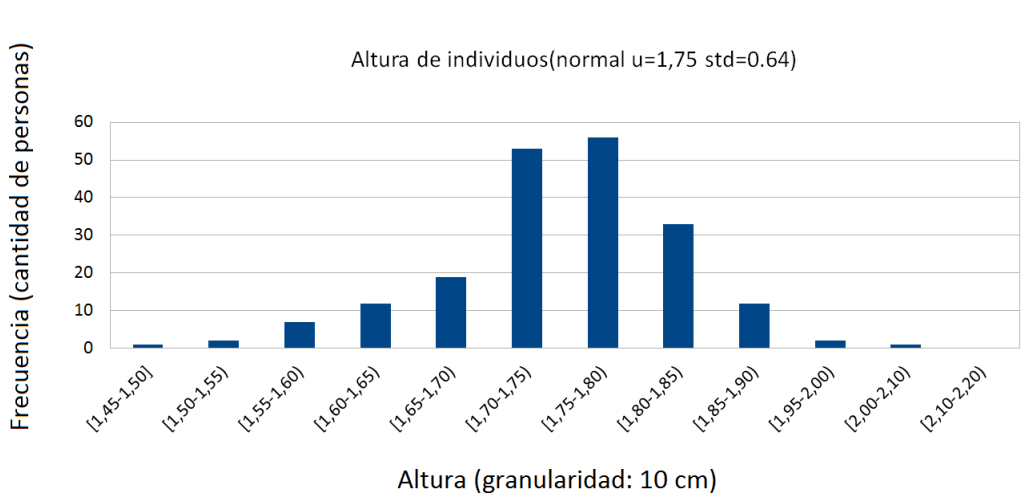

Figura 1: Histograma altura

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 coeficiente intelectual de las personas, conocido como IQ. Según el siguiente ranking, vemos que una inteligencia normal debería estar 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	Límite
69 o menos	Extremadamente bajo

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

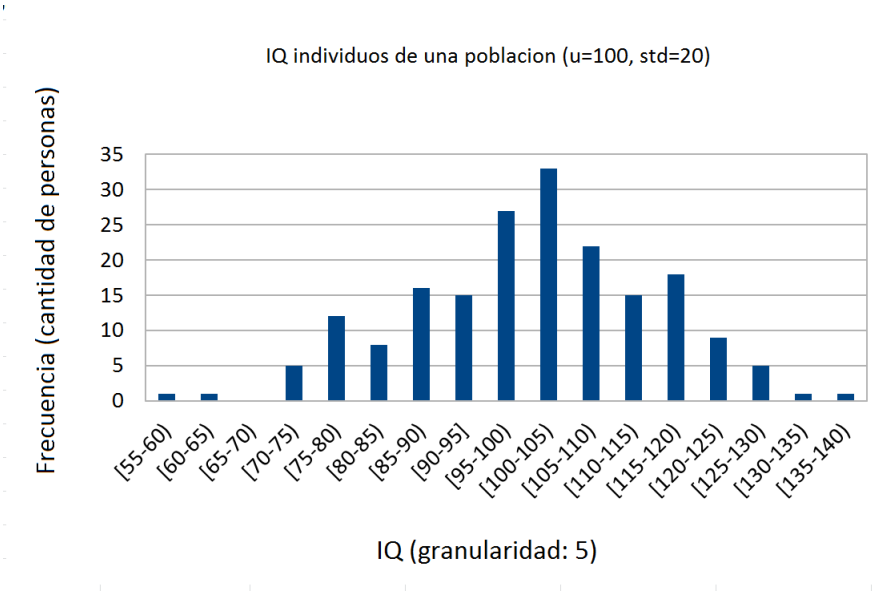

Figura 2: Histograma IQ

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. Distribución uniforme (Ejemplos)

La distribucion uniforme es una de las mas conocidas, y como la nomal, una de las mas presentes en la realidad. Esta distribución es muy simple. Básicamente plantea que 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 clásico para entender esta distribución, es el lanzamiento de un dado de 6 caras (no cargado), en donde $S = \{1, 2, 3, 4, 5, 6\}$ y cada elemento tiene la misma probabilidad de salir, osea $\frac{1}{6}$.

Tiempo de espera de un colectivo

Otro ejemplo un poco más interesante, es si tomamos un rango de tiempo, y medimos el tiempo de espera de un colectivo en ese rango. Si bien este ejemplo dependerá mucho sobre que línea de colectivo se haga el muestreo, se puede tomar un rango en particular en el cual sabemos que el tiempo de espera no será mayor o menor a eso. A continuación, se presenta un histograma sobre un *dataset* sobre los tiempos de cierto colectivo en el rango de 5 minutos a 30 minutos.

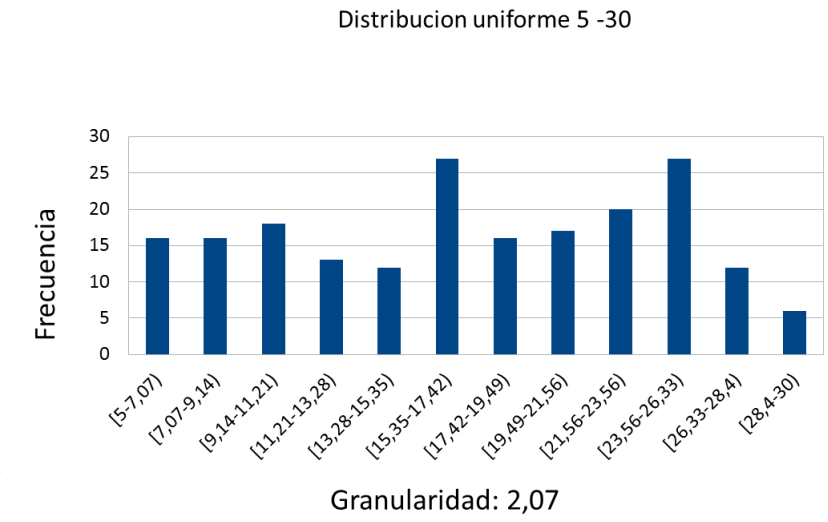

Figura 3: Histograma tiempos de espera de colectivo

3.2. Análisis de los estimadores: Parámetro fijo

3.2.1. Distribución normal

Caso 1

Parametro	20
Columna	C2
Valor maximo	1002
Valor minimo	-671
Distribución	Normal, Media=200
Selectividad de	Igualdad

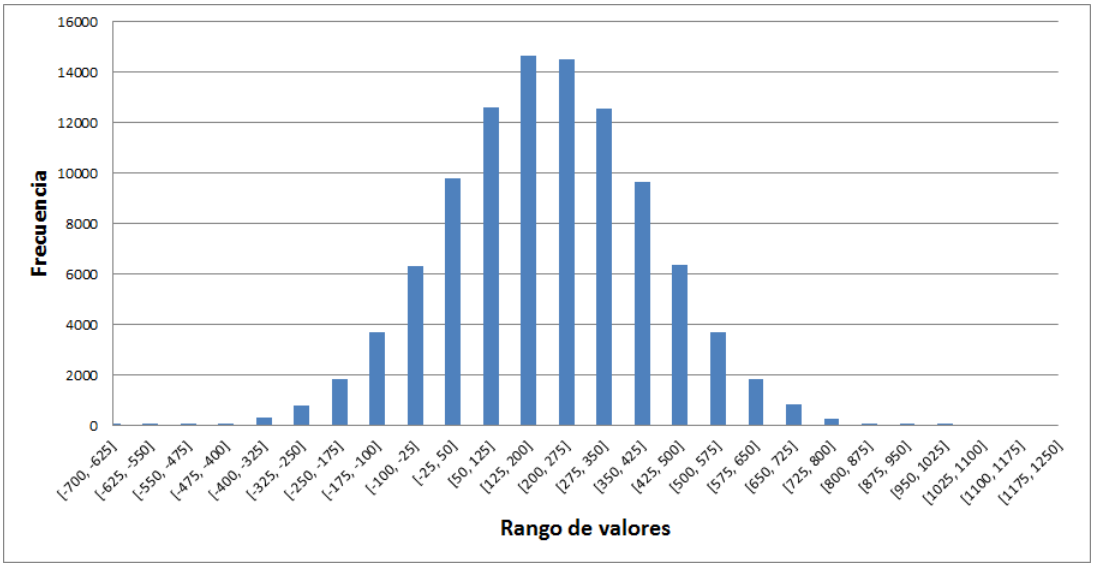

Figura 4: Gráfico de distribución de la columna C2 del set de datos de la cátedra

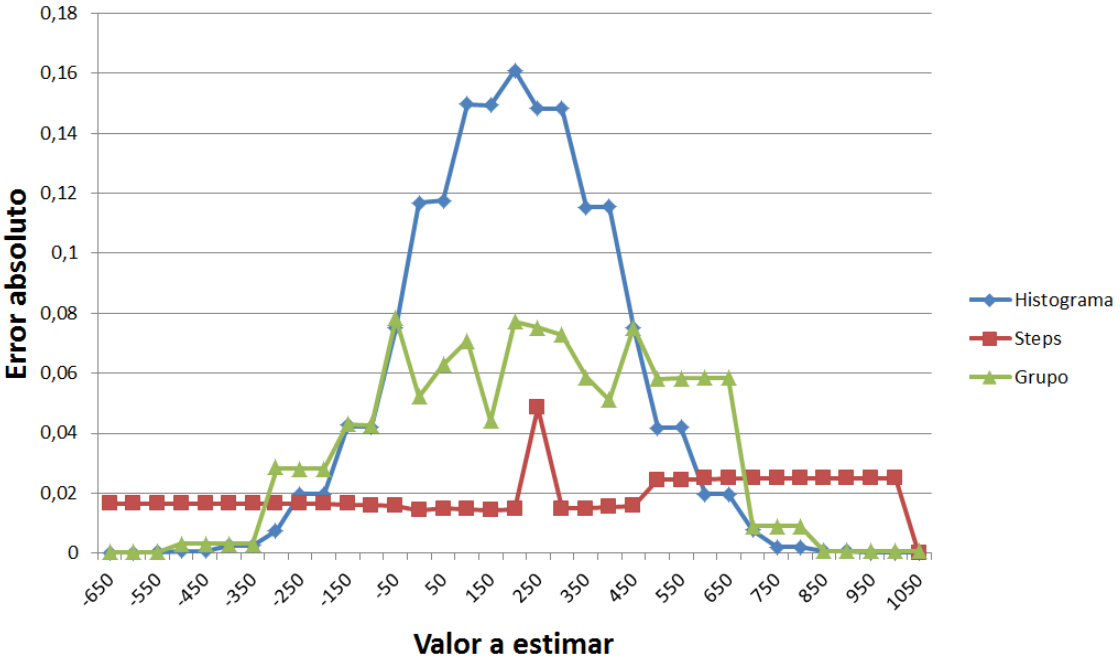

Figura 5: Errores variando valor a estimar con parámetro fijo

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 gráfico de la figura 5 se ve como, teniendo los parámetros 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 el Classic Histogram se comporta mejor en los casos que están por afuera del desvío 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 cercano a la media. En la media, se ve como el error llega al máximo.

En cuanto al estimador ideado por el grupo, se ve como al igual que el Histograma Clásico, se obtiene errores muy chicos en valores lejos del desvío standard de la normal. 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 del Grupo, es un histograma clásico pero con una re-distribución en los *bins* con más valores, los cuales en este caso estarán cerca de la media.

Caso 2

Parametro	20
Columna	C2
Valor maximo	1002
Valor minimo	-671
Distribución	Normal, Media=200
Selectividad de	Mayor

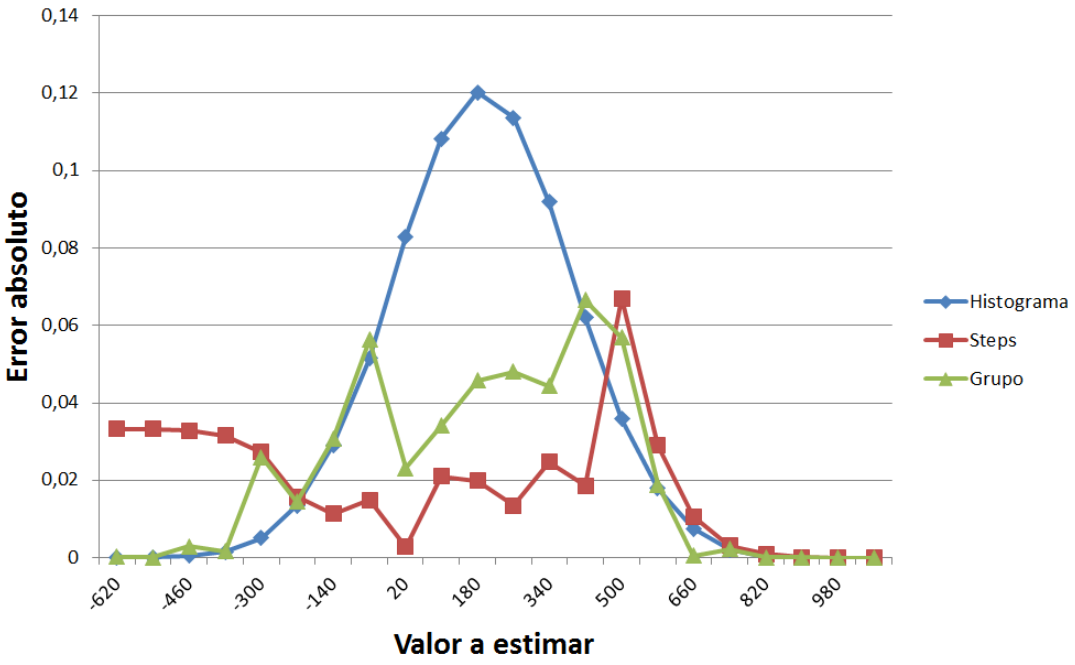

Figura 6: Errores variando valor a estimar por mayor con parámetro fijo

Usando el mismo set de datos uniforme que se uso anteriormente, vemos en la figura 9 el error de la selectividad pero estimando por mayor. En este caso sucede algo muy parecido a lo que sucede al estimar por igual, con la salvedad de que el de pasos distribuidos no resulta tan constante durante todo el rango.

El error del histograma clasico de nuevo es cada ves mayor a medida que se acerca a la media de la normal.

El estimador del grupo en esta ocasión resulto ser casi tan bueno como el de pasos distribuidos. Al parecer tiene una buena performance en los datos con distribución normal.

3.2.2. Distribución uniforme

Caso 3

Parametro	20
Columna	C0
Valor minimo	0
Valor maximo	999
Distribución	Uniforme, Media = 4950
Selectividad de	Igualdad

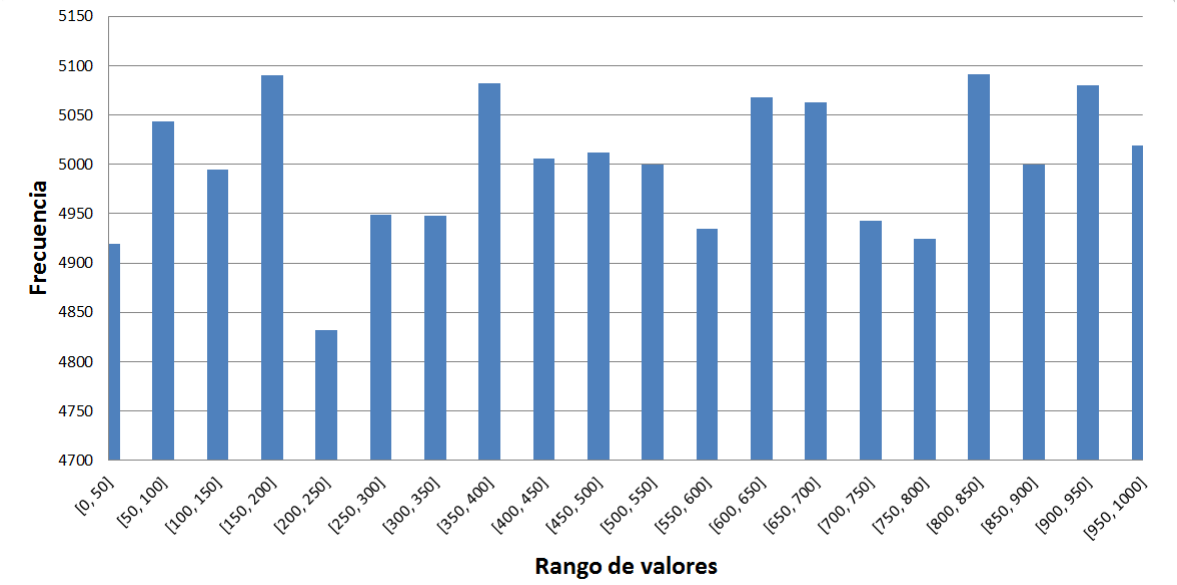

Figura 7: Gráfico de distribución de la columna C0 del set de datos de la cátedra

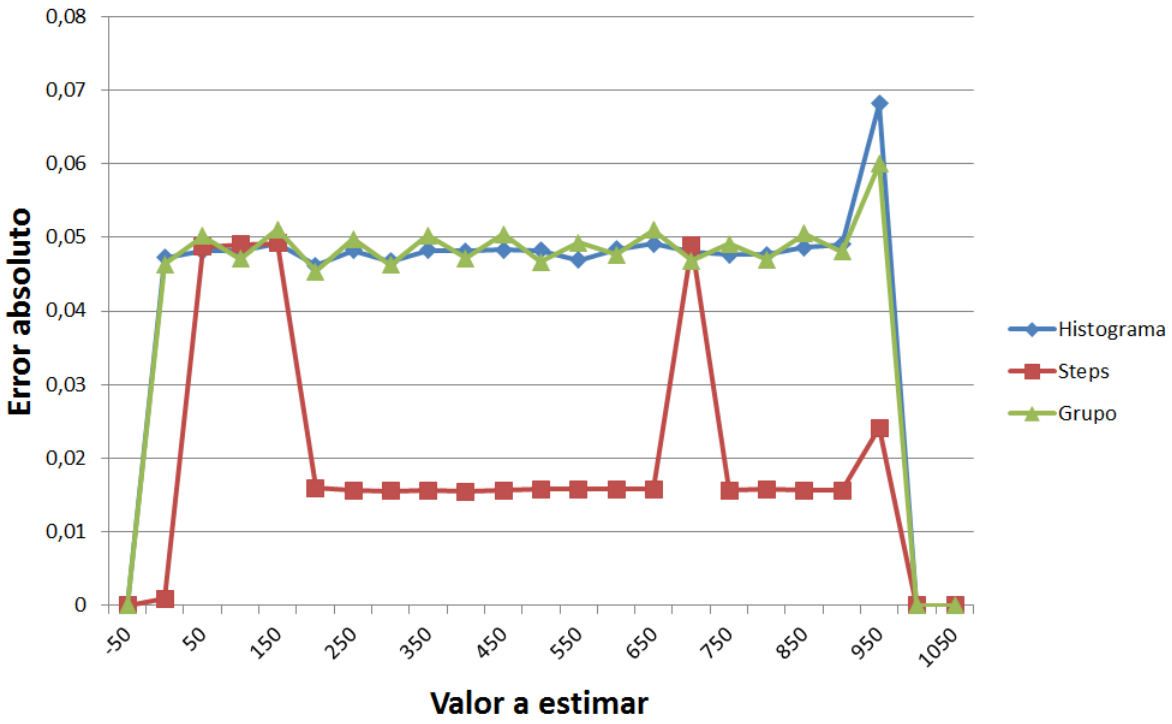

Figura 8: Errores variando valor a estimar con parámetro fijo

En la figura 7 se ve como la distribución de los datos esta ves es una Uniforme que varía al rededor de 4950.

Según se ve en la figura 8, no hay una gran diferencia esta vez con el Histograma Clásico y el estimador implementado por el grupo.

En este caso, se puede apreciar como claramente "Pasos Distribuidos" obtiene errores bastante menores en casi todo el rango.

Según pareciese, en los valores donde la distribución uniforme toma valores altos, los errores de pasos distribuidos aumentan bastante, igualando a los obtenidos en los otros 2 estimadores.

Caso 4

Parametro	20
Columna	C0
Valor minimo	0
Valor maximo	999
Distribución	Uniforme, Media = 4950
Selectividad de	Mayor

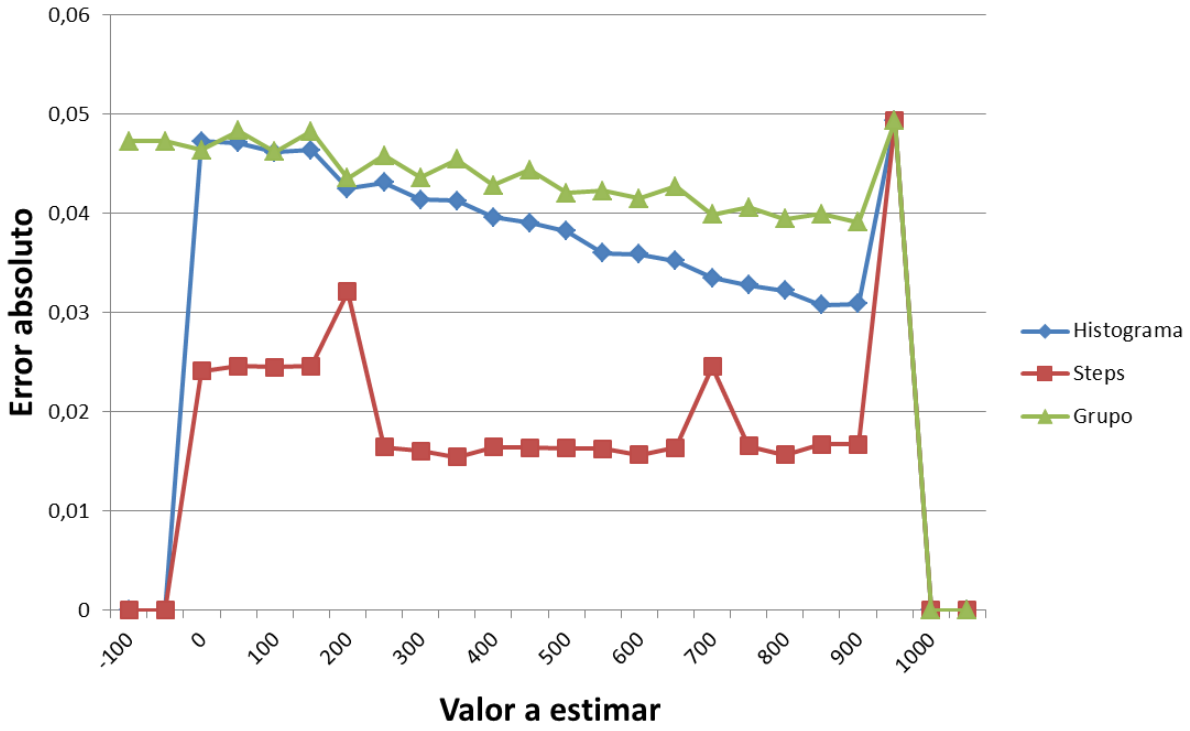

Figura 9: Errores variando valor a estimar por mayor con parámetro fijo

En la figura 9 se realizó un test igual al caso 2, pero estimando por Mayor en vez de por igual. Ya se puede ver como la distribución es un factor muy importante al momento de utilizar los estimadores.

En este caso, el estimador del grupo fue incluso peor que el histograma clásico, y se ve como la diferencia de error aumenta a medida que se acerca al valor más alto del rango.

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 lo llaman *steps*.

Se realizaron las comparaciones con un método exacto calculado sobre la colección de datos testeados.

3.3.1. Distribución uniforme

- **Operación por igualdad**

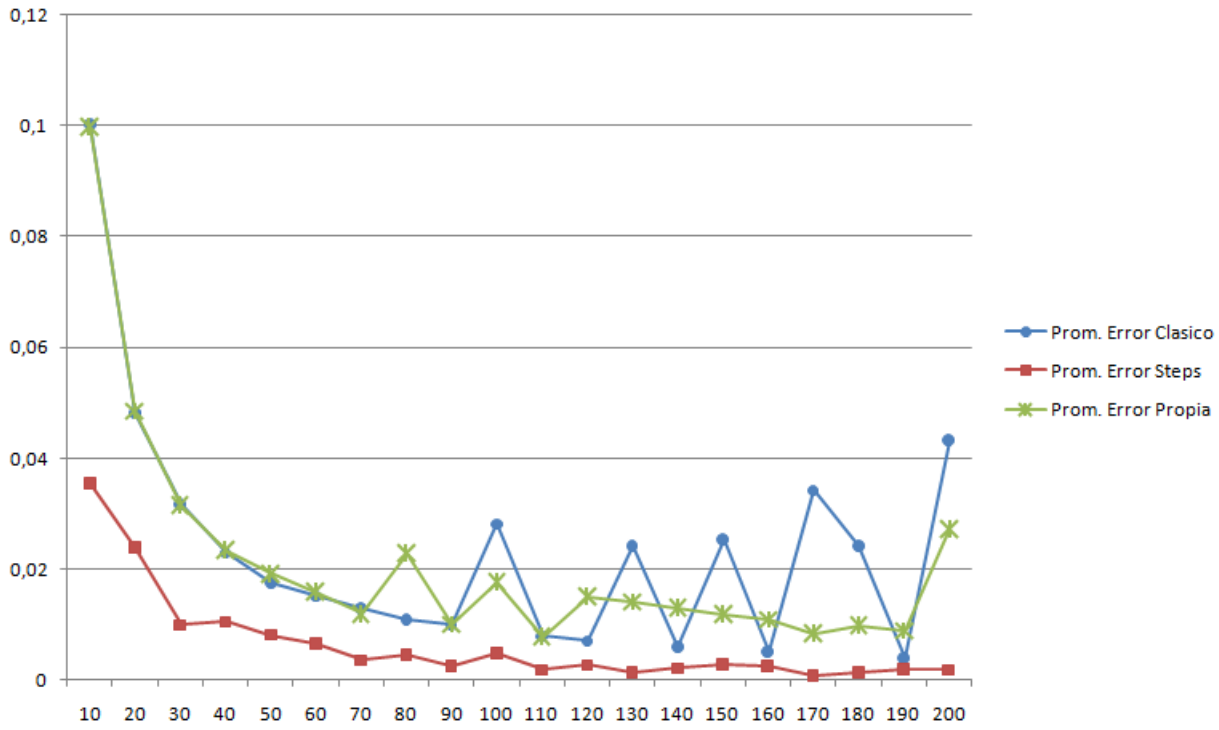

Figura 10: Error promedio de la Columna C0 de la tabla brindada por la materia

En la figura, se ve claramente como aumentando la cantidad de buckets el error disminuye. Esto se debe a que, al aumentar la cantidad de buckets, se aumenta la granuralidad sobre el rango de valores. Se puede apreciar como para el histograma clásico y el estimador de grupo afecta todavía mucho más que con el de pasos distribuidos. En éste último, al principio mejora bastante (hasta los 70 buckets) pero se estabiliza en un cierto valor y deja

de mejorar. En cambio. con los otros dos va mejorando pero se puede ver como se vuelve inestable para ciertos valores del parámetro. Creemos que se debe a ciertos casos bordes con respecto a la cantidad de elementos y los valores presentes de un rango y la cantidad de buckets.

- **Operación por mayor**

Ahora los estimadores los mediremos por su selectividad en un rango buscando ciertos valores en las distribuciones fijos y variando el parámetro.

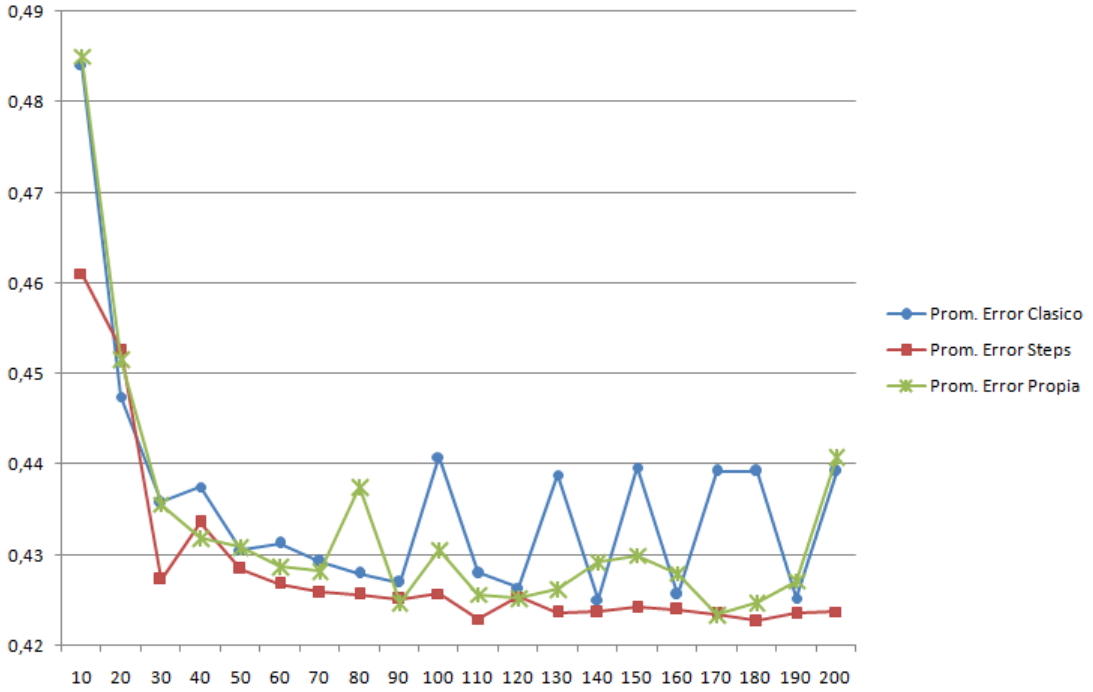

Figura 11: Error promedio de la Columna C0 de la tabla brindada por la materia

Podemos apreciar la similitudes entre este gráfico y el de la Figura 10. Pero los valores de error son mucho mayores. Ésto se debe a que se acumula error, que si bien puede pasar que entre errores se cancelen, creemos que en este caso se acumulan. A pesar de esto, consideramos correcto que mantengan la *forma* ambos gráficos.

Al igual que en el gráfico anterior, notamos como el histograma clásico y el estimador propio son inestables. Además, vuelve a pasar lo mismo con el de pasos distribuidos, mejora hasta cierto punto donde luego de determinado valor del parámetro tiende a un valor fijo.

De estos gráficos podemos concluir que para las distribuciones uniformes, el que *mejor* se comporta es el estimador de pasos distribuidos debido a su estabilidad y menor valor de error.

3.3.2. Distribución normal

- **Operación por igualdad**

Usando la misma metodología de tests, analisaremos para colección de datos con distribución normal.

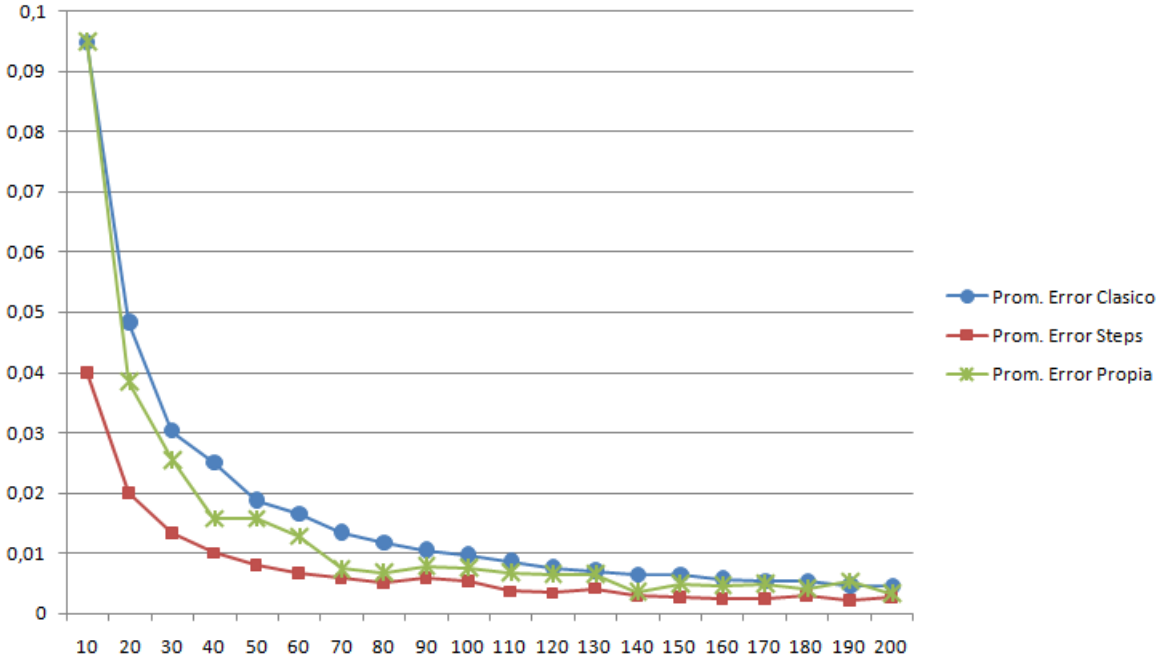

Figura 12: Error promedio de la Columna C2 de la tabla brindada por la materia

Nuevamente, todos mejoran a medida de que se aumenta la cantidad de buckets. Pero, en este caso, los tres son estables y, llamativamente, convergen hacia un mismo valor de error. Es decir, que a partir de cierto punto, no hay diferencia apreciable en cuanto al error entre los tres estimadores. Particularmente, cuando la cantidad de buckets es igual a 200, por el gráfico se ve que la diferencia entre los estimadores es menor a 0,0025 aproximadamente.

- **Operación por mayor**

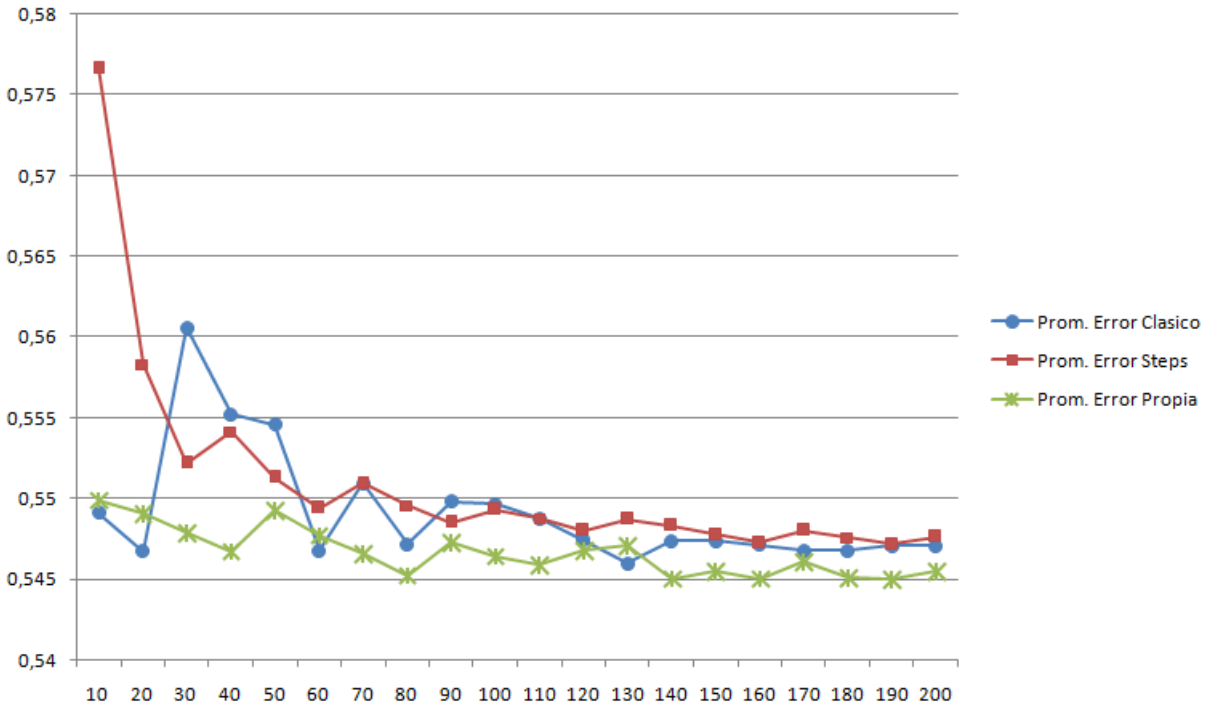

Figura 13: Error promedio de la Columna C2 de la tabla brindada por la materia

En esta figura, vemos como el que empieza con mayor error es el de pasos distribuidos y al igual que los otros dos va disminuyendo el error a medida que se incrementa la cantidad de buckets. Con respecto al de histograma clásico, empieza inestable y luego se estabiliza, ocurría lo contrario en los gráficos 10 y 11. En cuanto al estimador propio, vemos que no es tan significativa la mejora comparándola con los otros dos pero es el que menor error genera.

Se ve que, en este caso, los resultados dieron muy diferentes a todos los anteriores gráficos. Mostrando como, los estimadores dependen del tipo de distribución y, además, del tipo de operación que se realice.

3.4. Análisis para distintas columnas con parámetro variable

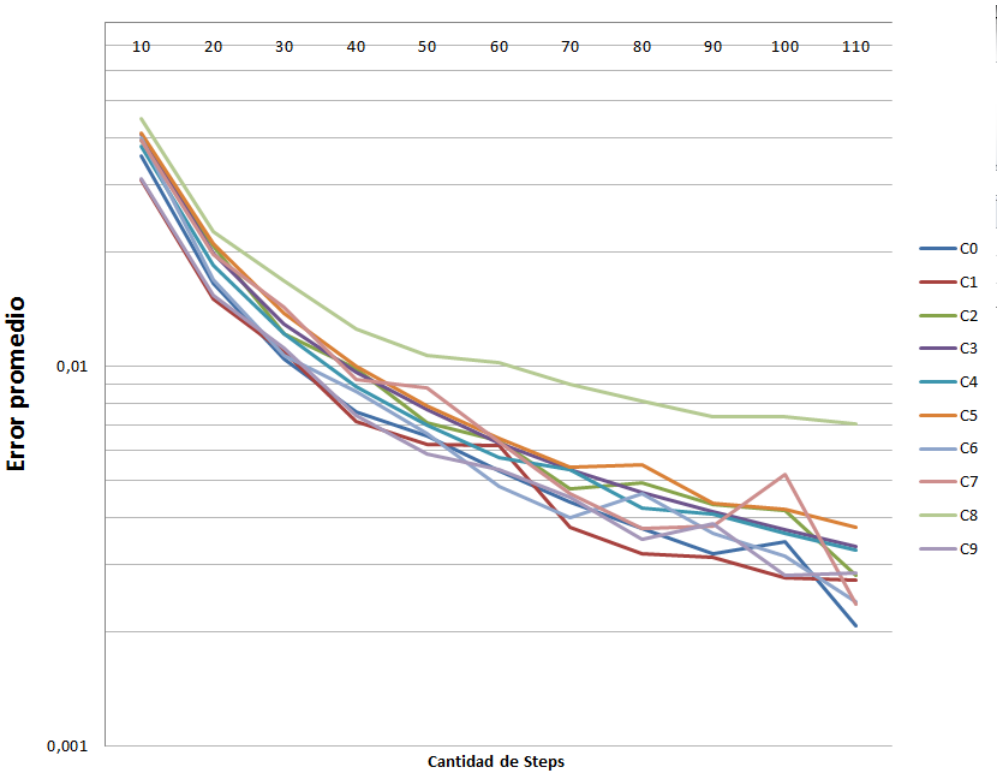

Figura 14: Error promedio en Steps para todas las columnas de la tabla brindada por la materia

A continuación, realizamos un análisis del estimador "Distributed Steps", el cual obtuvo los mejores resultados según nuestros análisis previos.

En la figura 14 se ve un gráfico que se realizo utilizando los errores promedio para cada columna, variando a la vez la cantidad de Steps del estimador y estimando por igualdad. Como error promedio consideramos al promedio de los errores obtenidos para los distintas estimaciones de los valores de una misma columna. Una vez calculado el error promedio para ciertos Steps y una columna en particular, repetimos el proceso variando la cantidad de Steps para todas las columnas. Se ve como el error promedio de cada columna, es muy similar para pocos Steps. Pero al aumentar la cantidad de steps, mejora mucho el error. Esta mejora es independiente de la distribución de la columna, ya que la diferencia de errores entre columnas, es casi despreciable.

Solo para la columna C8 se ve como el error es considerablemente mayor al resto. Esta columna es particularmente distinta a las demás, ya que sus datos van de 0 a 41 con una distribución que no es ni normal ni uniforme. Pareciese como una "media campana" con media en el 0, y que va disminuyendo a medida que se acerca al 41, en donde pasa a ser 0. En la archivo "/tests/Determinar Distribuciones/c8_distro.xlsx" se puede ver el gráfico del mismo. Sin embargo, por más de que en esta columna el error haya subido un poco, sigue estando en un valor bastante bajo, por lo que se puede concluir que "Distributed Steps" la distribución de los datos no influye demasiado en el error.

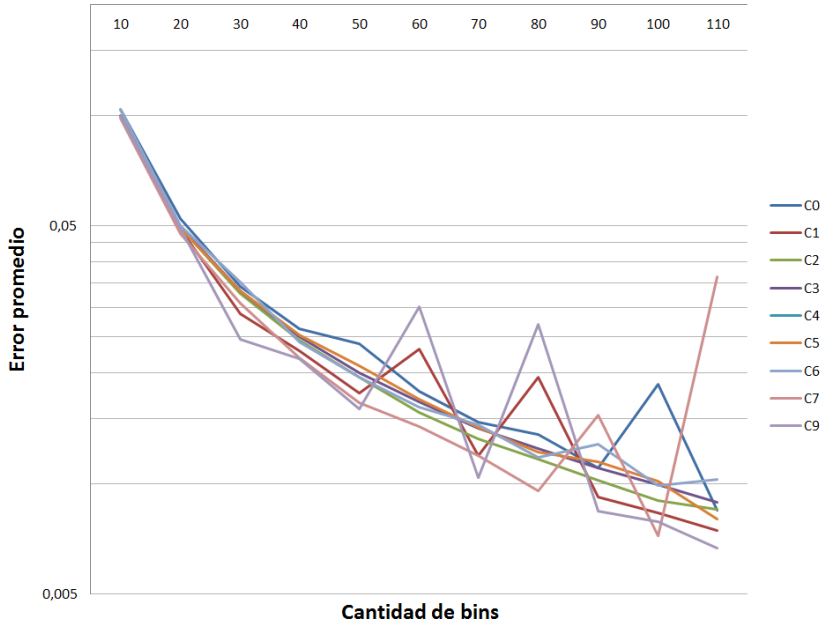

Figura 15: Error promedio en Histograma Clásico para todas las columnas de la tabla brindada por la materia

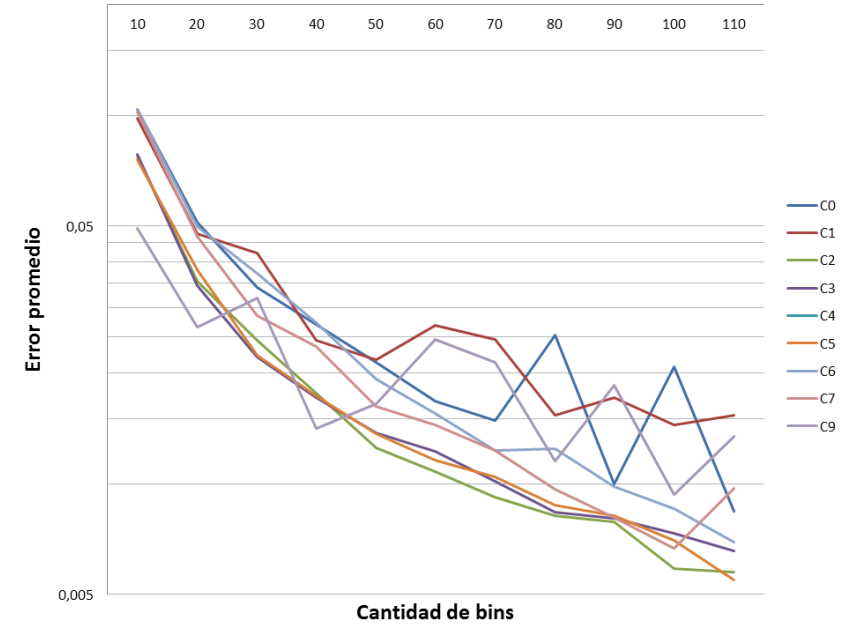

Figura 16: Error promedio en Estimador del grupo para todas las columnas de la tabla brindada por la materia

Se puede ver en las figuras 15 y 16 el mismo análisis realizado para Pasos Distribuidos, pero esta vez para Histograma clásico y el estimador hecho por el grupo respectivamente. En estos gráficos no se incluyó la columna C8, ya que la misma tiene datos que del 0 al 41, por lo que tienen un rango de 42 elementos, y para 50 buckets se tiene 1 bucket para cada valor, por lo que el error pasa a ser 0.

Comparándolos con el gráfico de Steps, a primera vista, se ve como en general Steps se comporta mejor considerando un valor del parámetro fijo.

A diferencia del de Pasos Distribuidos, el histograma clásico no es tan independiente de la distribución de los datos. Se puede ver como para las distintas columnas el error varía bastante.

En cuanto al estimador del grupo, lo que se puede apreciar es que para algunas columnas, se comporta mejor que el Histograma Clásico, y difícilmente se comporta mejor que Steps. Debido a lo analizado anteriormente, sabemos que el estimador del grupo tenía una mejor performance sobre el Clásico solo en las distribuciones normales. En las uniformes, los errores no variaban mucho entre uno y otro. Las columnas donde la performance es mejor que en el histograma, son las C2, C3 y C5. En el anexo a este informe, en la carpeta "/tests/Determinar Distribuciones" se encuentran los gráficos de las columnas mencionadas. No fueron incluidos en este informe, debido a que eran demasiados quedando demasiados gráficos para mostrar. También, se encuentran graficadas las distribuciones de las columnas C1 y C9, que son las que obtuvieron mayores valores de error.

En esos gráficos, se puede ver como las distribuciones de esas columnas C2, C3 y C5 son todas normales, y C1 y C9 son de otra distribución muy distinta, que no es normal ni uniforme. Por lo que se puede comprobar que nuestro análisis previo, en donde nuestro estimador se comportaba mejor en distribuciones normales, es correcto.

3.5. Pasos distribuidos: medición del error

Para implementar el estimador según el paper de Piatetsky-Shapiro, nos basamos en la implementación que estima con el menor error para el peor caso. Hay otra implementación que se basa en los casos de error promedio más chico.

En su paper, Piatetsky-Shapiro determinan casos donde puede caer un valor en algún tipo de step. Siendo X el valor buscado ya sea por igualdad o por mayor, los casos y sus errores son los siguientes:

- **A) X está entre steps:** $\frac{2}{3} * S$
- **B1) X es igual al valor de un step I pero I no es el primer step ni el último:** $\frac{1}{S}$
- **B2) X es igual al valor de varios steps pero no incluye al primer step ni al último:** $\frac{1}{S}$
- **C) X es igual al valor de uno o varios steps, pudiendo incluir al primer o último step:** $\frac{1}{S}$
- **D) X está afuera del rango de posible valores:** 0

Aclaración: En el paper no figura cota de error para el caso C, por lo que asumimos que es igual a los anteriores, ya que la diferencia está en que puede incluir al primer step o al último.

A continuación presentamos el gráfico derivado de la figura 14:

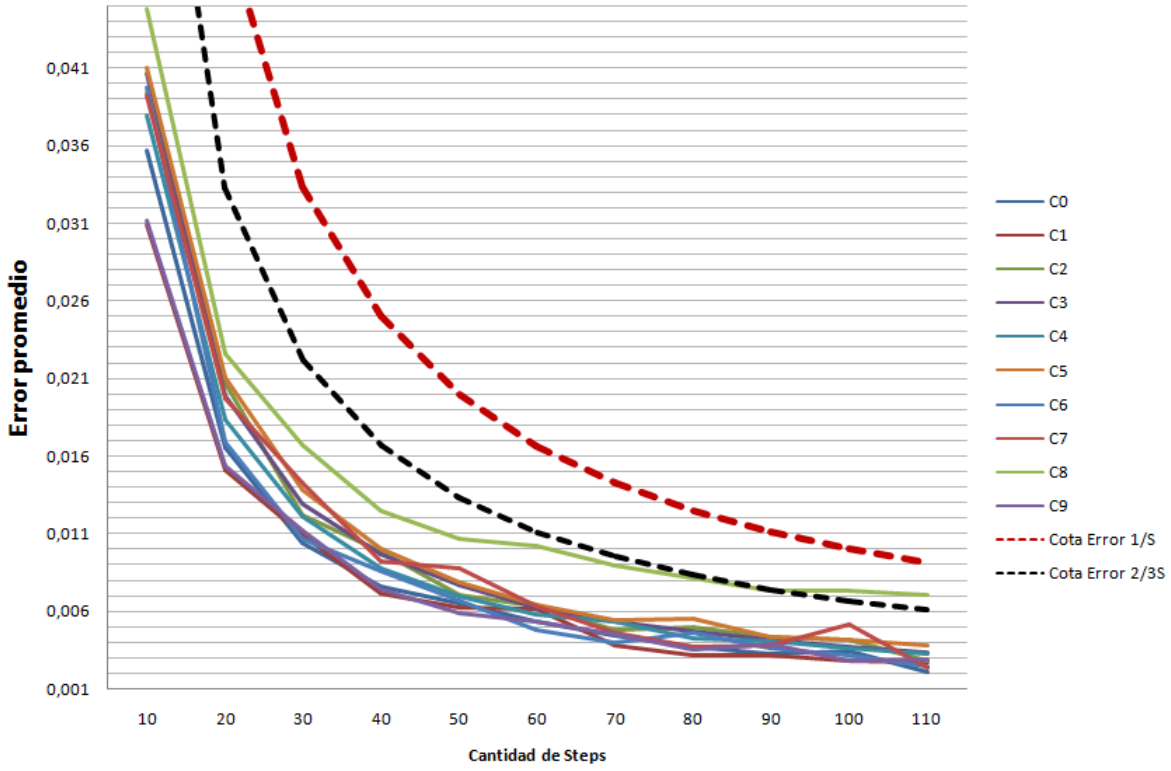

Figura 17: Error del estimador de pasos distribuidos sobre todas las columnas y las cotas de error

Como podemos ver, todas salvo la columna C8, cumplen con las cotas. Se observa, que a medida que varía la cantidad de steps, no cambia considerablemente la diferencia entre el error y las cotas.

En cuanto a la columna C8, analizando su distribución y el algoritmo de pasos distribuidos pudimos comprobar que hay un problema con el algoritmo del cual no se habla en el paper.

El rango de la columna C8 es de 0 a 41. El algoritmo de pasos distribuidos determina el valor de un step I como el elemento que se encuentran en la posición (1 + I*N), donde $N = \frac{T-1}{S}$ y S es el valor del parámetros que simboliza la cantidad total deseada de steps y T la cantidad total de tuplas en la tabla. Debido a esto, cuando la cantidad de steps es mayor a la cantidad de elementos por steps se produce un error. Debido al rango de las demás columnas, nuestros testeos los realizamos con un valor alto de cantidad de steps. Alto comparando con el espectro de la columna C8. Adjudicamos a eso el por qué a partir de cierto punto el error rompe la cota para esa columna.

4. Discusión y conclusiones

4.1. Discusiones

Según los análisis realizados, pudimos comparar el rendimiento de los tests en cuanto a los errores que se producen al estimar para las distintas distribuciones. Según parámetros de los estimadores, valores estimar; estimando por igualdad o por mayor.

Lo que sacamos en claro es que, independientemente de la distribución de los datos, el estimador "Distribution Steps" mantiene un error constante para todos los valores del rango de los datos. En cambio, "Classic Histogram", en distribuciones normales, se comporta muy bien (mejor que Distribution Steps) en valores alejados de el desvio standard, pero muy mal para valores alrededor de la media.

Para el estimador realizado por nosotros, vimos que para distribuciones normales se comporta igual que Classic Histogram en valores por afuera del desvío standard, y mejor que el mismo para valores cercanos a la media. Sin embargo, "Distribution Steps" se comporta mejor para estos últimos casos.

Para distribuciones uniformes, vimos que en algunos casos el error promedio en "Classic Histogram" varia mucho al variar el parámetro, al igual que el estimador propio. Incluso, vimos algunos casos en donde el error promedio de "Classic Histogram" empeora al aumentar el parámetro. Como por ejemplo, se puede ver en los gráficos del as figuras 10, 11 y 15.

Sin duda, el más afectado por la distribución de los datos, fue el estimador propio. Se comporta bastante bien para casos en distribuciones normales, como se puede apreciar en los gráficos de las figuras 16 (para este gráfico ver tambien gráficos con distribuciones de las columnas: carpeta "/tests/Determinar Distribuciones") y 13. En esta última se puede ver como el estimador propio resultó mejor que "Distribution Steps".

4.2. Conclusiones

Según nuestros análisis, concluimos que "Distribution Steps" puede resultar un buen estimador cuando no se conoce la distribución de los datos, ya que de esta forma se puede acotar el error de estimación de la selectividad a valores bastante bajos y sobre todo constantes para todos los datos.

También pudimos ver como realizando una pequeña mejora a un Histograma Clásico, se pueden obtener resultados interesantes para distribuciones normales, y no tanto para otro tipo de distribuciones.

Por ultimo, también vimos como hay casos en donde aumentar el parámetro demasiado, puede no afectar en nada en los errores, y incluso en los estimadores basados en Histogramas, puede hasta empeorar el error y encima aumentar el tiempo de construcción de los mismos. Por lo que siempre es conveniente encontrar un valor de parámetro balanceado, que disminuya el error pero que no haga q los tiempos de construcción del estimador no sea demasiado alto.

5. Aclaraciones

Ajunto a este informe, se encuentran las carpetas:

- **src**: Contiene el código correspondiente a los estimadores, así como el *interfaz_bd.py* que se pedia en el enunciado.
- **tests**: Contiene los archivos Excel donde se encuentran los gráficos utilizados en todo el informe. No se provee un script para generarlos como se pedía para poder variar su datos, pero en los Excel se puede modificar el set de datos principal, y los gráficos se redibujarán con esos datos.
- **/tests/Determinar Distribuciones**: Contiene los gráficos de las distribuciones del set de datos provisto por la materia (no todas, solo las que necesitamos para realizar los análisis).