

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	3
3.1. Distribuciones utilizadas	3
3.1.1. Distribucion normal (Ejemplos)	3
3.1.2. Distribucion uniforme (Ejemplos)	5
3.2. Análisis de los estimadores	6

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 que el motor realiza. Por eso mismo, se trata de minimizar estas operaciones.

Una de las formas de poder minimizar estas operaciones, es el poder conocer aproximadamente cual puede ser la distribución de un set de datos, y de esta forma poder estimar cuantas tuplas se obtendrán por el echo de aplicar un filtro (WHERE) u otro. El echo 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 minimice.

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

Aca no se si habria que explicar cada estimador.. tecnicamente no lo pide el enunciado.. asi q esto se podria obviar :P, si hay q explicar mas adelante el estimadir propio

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 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.

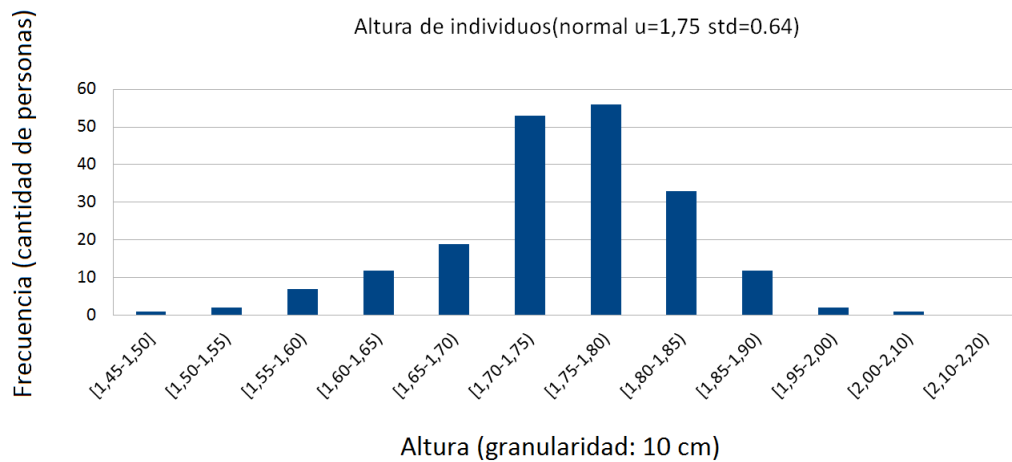


Figura 1: Histograma altura

Se puede ver como para el muestreo realizado, la mayoría 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 estar entre los 90 y los 109 de coeficiente intelectual, por lo que es de esperar que la mayor parte de la población esté 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.

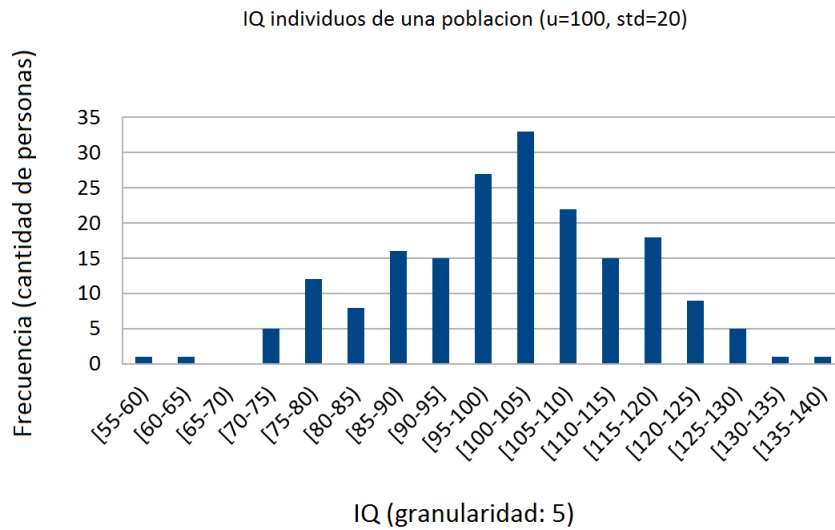


Figura 2: Histograma IQ

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

3.1.2. Distribucion uniforme (Ejemplos)

La distribución uniforme es una de las más conocidas, y como la normal, una de las más presentes en la realidad. Esta distribución es muy simple, básicamente plantea que si tenemos un espacio muestral $S = \{m_1, m_2, \dots, m_n\}$, ($\forall m_i \in S, P(m_i) = 1/n$), o sea, todos los sucesos tienen la misma probabilidad de ocurrir.

El ejemplo más clásico para entender esa 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, o sea $1/6$.

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 de tiempo. Si bien este ejemplo dependerá mucho de sobre qué línea de colectivo hagamos 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.

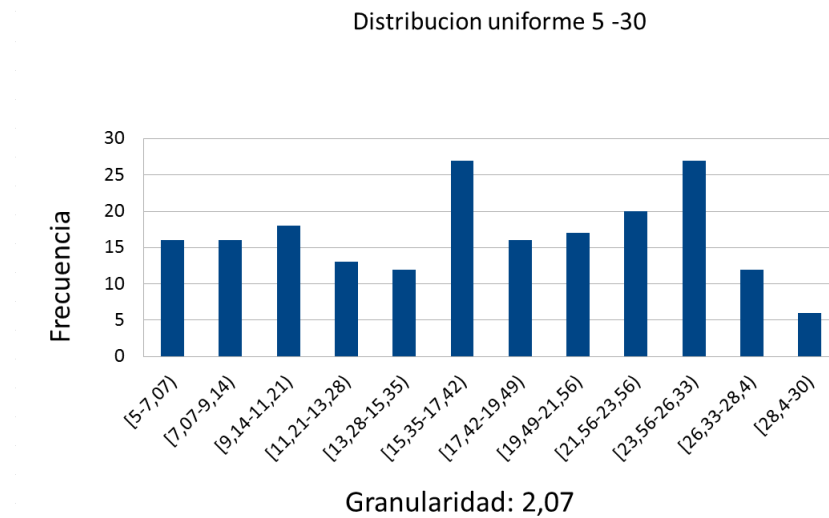


Figura 3: Histograma tiempos de espera de colectivo

3.2. Análisis de los estimadores

Primero realizaremos analisis sobre, un mismo set de datos, y un parametro 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 metodo y cuales son las ventajas y desventajas de cada uno.

Primero, intuitivamente podriamos 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 parametro 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 graficos comparando los 3 estimadores implementados.

Caso 1

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

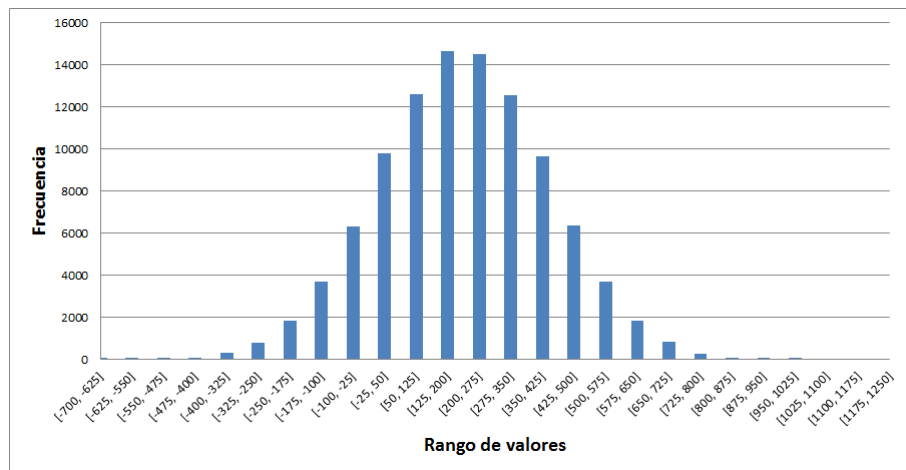


Figura 4: Grafico de distribucion de la columna C2 del set de datos del a caterda

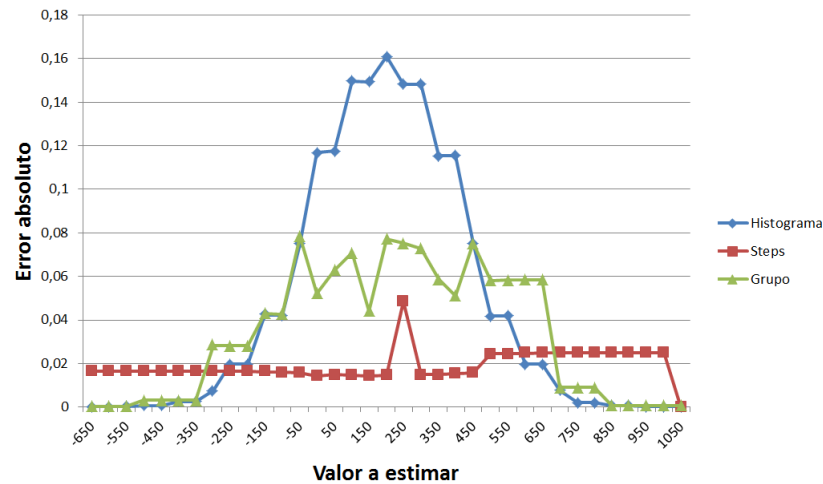


Figura 5: Errores variando valor a estimar con parametro fijo

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 5 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 vez, 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 estarían cerca de la media.