Histogramas

Guarda los valores más frecuentes

- Guardan la frecuencia de ciertos valores
  - Podría ser todos, pero ocupa más
- Las fórmulas se pueden adaptar para estimar con este conocimiento
  - Para la selección, si se conoce la cantidad exacta del valor buscado, se la usa
  - Para el Join, hay que adaptar la fórmula
- PostgreSql tiene la vista pg_stats con información
  - select * from pg_stats WHERE tablename = 'player'

Histogramas - Ejemplo

A	100	200	300	400	Otros
R.A	50	100	50		100
S.A	100		200	150	550

- Se guardan la cantidad de filas para los 3 valores más frecuentes para R y S
  - Pueden no coincidir cuales son los más frecuentes!
- Además se debe almacenar cuantos valores tiene el atributo en cada tabla:
  - V(A,R) = 7
  - V(A,S) = 14
- Del histograma surgen la cantidad de filas
  - n(R) = 300
  - n(S) = 1000

Histogramas - Ejemplo

A	100	200	300	400	Otros
R.A	50	100	50		100
S.A	100		200	150	550

- Para selección en R
  - Puedo tener el valor certero
    - Si la condición es A = 100, se devuelven 50 filas para R
  - Si no, mejor estimación
    - Si la condición es A = 500, hacer 100 / 4
    - V(A,R) = 7, pero ya se la cantidad de 3 valores entonces el “Otros” representa los otros 4 valores

Histogramas - Ejemplo

A	100	200	300	400	Otros
R.A	50	100	50		100
S.A	100		200	150	550

- Para join, un poco más complejo
- Sin tener en cuenta histograma:
  - Cantidad: $\frac{300 * 1000}{máx(7,14)} = 21{,}429$

Histogramas - Ejemplo

A	100	200	300	400	Otros
R.A	50	100	50		100
S.A	100		200	150	550

- Teniendo en cuenta el histograma
  - Para el valor 100 se combinan 50 * 100
  - Para el 300 se combinan 50 * 200
  - El resto de los valores, no sabemos con exactitud
    - Ojo! El valor de R.A = 200 y el de S.A = 400 pueden combinarse
      - Que no se sepa su cantidad exacta, no implica que no estén incluidos en “Otros”
      - Se estiman asumiendo que están en la tabla y tienen distribución equitativa

Histogramas - Ejemplo

A	100	200	300	400	Otros
R.A	50	100	50		100
S.A	100		200	150	550

- Para R.A = 400
  - 100 registros comparten los 4 “Otros” valores
    - Conocemos 3 de los 7 valores posibles
  - Se asumen 25 con valor 400, 75 con Otros
- Para R.S = 200
  - 550 registros comparten los 11 “Otros” valores
  - Se asumen 50 con valor 200 y 500 para otros

A	100	200	300	400	Otros
R.A	50	100	50	**25**	75
S.A	100	50	200	150	500

Histogramas - Ejemplo

A	100	200	300	400	Otros
R.A	50	100	50	25	75
S.A	100	50	200	150	500

- Se usan los nuevos valores para estimar
  - Para “Otros”, no considerar los 4 valores conocidos en V(A,R) y V(A,S)
- Cantidad: 50 * 100 + 50 * 100 + 50* 200 + 25*150 + $\frac{75 * 500}{máx(3,10)}$
  - 5,000 + 5,000 + 10,000 + 3,750 + 3,750
  - Cantidad del join: 27,500