Student ID: 1605029

HPDS Class Work, Week 2

Question 4-1

There are 40 nodes and 40,000 tuples. Here is how the tuples will be distributed:

$N_0 = [T_0, T_{40}, \dots, T_{39,960}]$

$N_1 = [T_1, T_{41}, \dots, T_{39,961}]$

$\vdots$

$N_{39} = [T_{39}, T_{79}, \dots, T_{39,999}]$

Question 4-2

There are 40 nodes and 40,000 tuples. He is how the tuples will be distributed:

a. $N_{h(NID_i)} = [T_i], \quad 0 \leq i < 40{,}000$

Explanation: We'll put the ith tuple in the node having id equal to the result of the hash of NID_i.

b. $N_{h(street_i, city_i, district_i)} = [T_i], \quad 0 \leq i < 40{,}000$

Question 4-3

a. Partition vector, ~~$V = [202,105,021, 202,105,041, \dots, 202,105,351]$~~

~~b. The required partitions are:~~

$V = [202105021, 202105041, \dots, 202105381]$

~~$P_0 = [202,105,001, \ldots, 202,105,020]$~~

b. The required partitions are:

$P_0 = [202105001, \ldots, 202105020]$

$P_1 = [202105021, \ldots, 202105040]$

$P_{19} = [202105381, \ldots, 202105400]$

Question 4-4

a. If there are n tuples in Person DB and m tuples in Parents DB, then with brute force technique we need to perform ~~(nm)~~ operations in the worst case. But with range partitioning, we can search the DB in the ~~four~~ following manner:

Search for all pair from $Partition_{1, Person}$ and $Partition_{1, Parents}$ ~~(nm)~~ ($\frac{nm}{16}$ operations in the worst case). Then for $Partition_{2, Person}$ and $Partition_{2, Parents}$ and so on.

b. For four nodes, the speed-up is four, since time elapsed has decreased four-fold. The scale-up depends on the problem size. If the problem size is unchanged, the scale-up is four-fold. But if the problem size is also increased as much as the time is decreased, the scale-up is one.

Question 5-1

a. Good, since all nodes have almost equal number of ~~tupe~~ tuples.

b. Good

c. Bad, since we wouldn't get any range from such partitioning.

Question 5-2

a. Good when the hash function is good and ~~partitioning~~ partitioning attributes form a key, since tuples will be equally distributed between nodes.

b. Good for point queries on partitioning attributes.

c. Bad, since for range queries, all nodes much be processed.

Question 5-3

a. Good, since it provides data clustering.

b. Good for point queries on partitioning attributes.

c. Good if the result tuples are from a few blocks.

Question 6-1

a. These skews can occur:

1. Attribute-value skew: Due to skew inherent in the dataset.

b. These skews can occur:

1. Attribute-value skew: ~~Due~~
2. Partition skew: When the partition vector is badly chosen.
3. Execution skew

Question 6-2

a. The type of the histogram is equi-width.

b. Total frequency $= 5(45+35+25+50+15) = 4 \times 212.5$

Partition vector, $V = [5, 10, 17]$

Partitions:

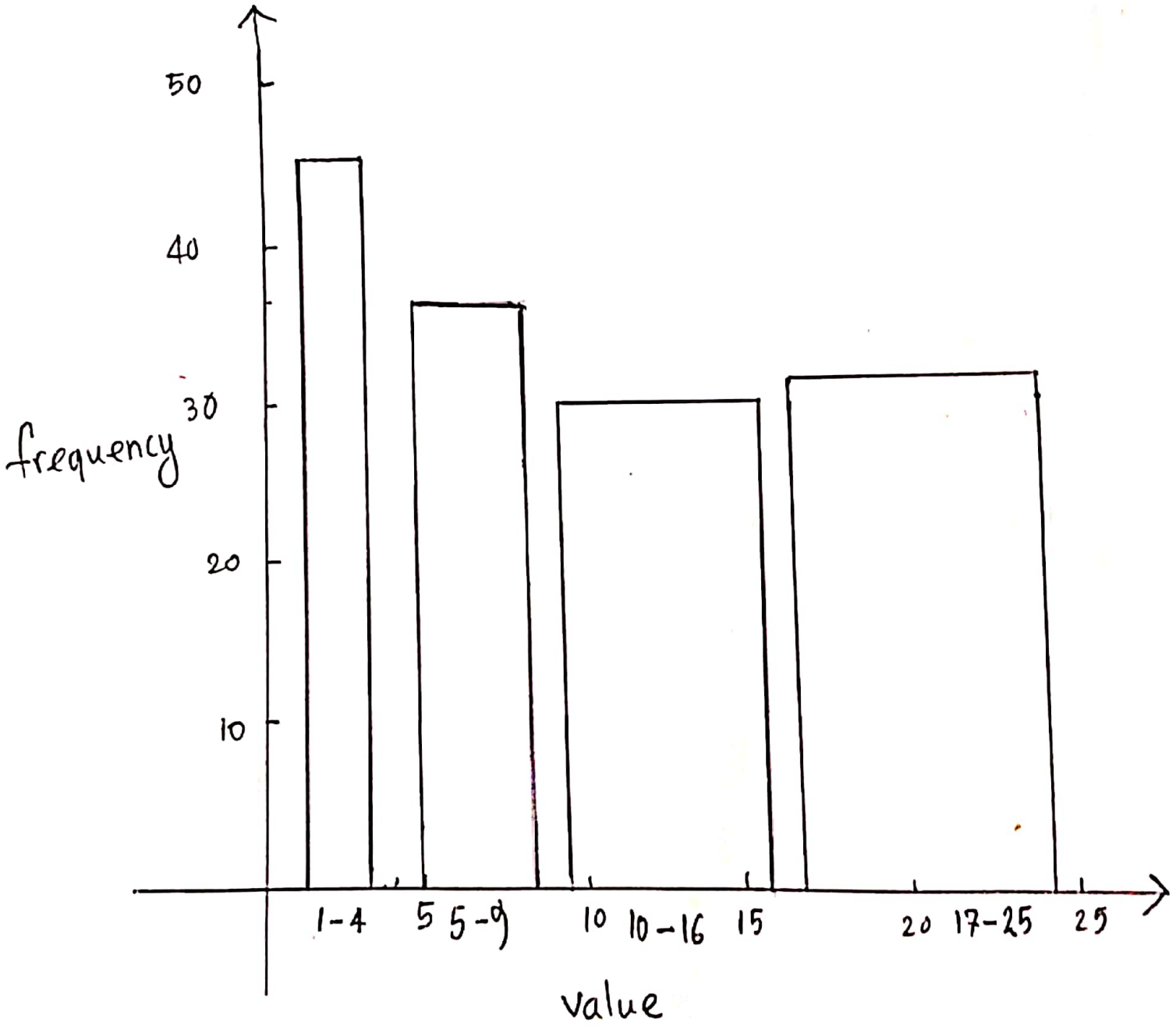
$P_0 = [1, 2, -3, 4]$ (Avg. Freq., $\bar{f} = 45$)

$P_1 = [5, 6, 7, 8, 9]$ ($\bar{f} = 37$)

$P_2 = [10, 11, 12, 13, 14, 15, 16]$ ($\bar{f} = 30$)

$P_3 = [17, 18, \ldots, 25]$ ($\bar{f} = 30.6$)

c. [Next Page]



Scanned with CamScanner