ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ
ΜΥΕ041 - ΔΙΑΧΕΙΡΙΣΗ ΣΥΝΘΕΤΩΝ ΔΕΔΟΜΕΝΩΝ
ΕΡΓΑΣΙΑ 1

Κίτσιος Κωνσταντίνος 4388

March 31, 2023

Contents

1 Πηγαίος κώδικας

Η εκτέλεση του προγράμματος στο τερματικό γίνεται με την εντολή:
python3 assignment1.py filename fieldname
Όπου filename το όνομα του αρχείου εισόδου και
fieldname το όνομα του πεδίου του αρχείου εισόδου (Θα πρέπει να έχει αριθμητικές τιμές).
Ο πλήρης κώδικας της εργασίας είναι ο εξής:

```
1  # Konstantinos Kitsios, AM: 4388
2  # How to execute:
3  # python3 assignment1.py <filename> <fieldname>
4
5  import csv
6  import sys
7
8  def equiwidth_histogram(data, bins):
9      min_value = min(data)
10     max_value = max(data)
11     bin_width = (max_value - min_value) / bins
12     bins_ranges = [min_value + i * bin_width for i in range(bins+1)]
13     hist_data = [0] * bins
14     for d in data:
15         bin_index = int((d - min_value) // bin_width)
16         if bin_index == bins:
17             bin_index -= 1
18         hist_data[bin_index] += 1
19     return hist_data, bins_ranges
20
21 def equidepth_histogram(data, bins):
22     sorted_data = sorted(data)
23     n = len(data)
24     bin_size = n // bins
25     remainder = n % bins
26
27     bins_ranges = []
28     hist_data = []
29     start = 0
30     for i in range(bins):
31         if i == bins - 1:
32             end = n
33         else:
34             end = start + bin_size
35             if i >= bins - remainder:
```

```
36                  end += 1
37
38          bins_ranges.append(sorted_data[start])
39          hist_data.append(end - start)
40
41          if i == bins - 1:
42              bins_ranges.append(sorted_data[end - 1])
43
44          start = end
45
46      return hist_data, bins_ranges
47
48  def estimate_tuples(hist_data, bins_ranges, a, b):
49      count = 0
50      for i in range(len(bins_ranges)-1):
51          if b <= bins_ranges[i]:
52              break
53          if a < bins_ranges[i+1]:
54              percentage = (min(b, bins_ranges[i+1]) - max(a, bins_ranges[i])) /
55                           (bins_ranges[i+1] - bins_ranges[i])
56              count += percentage * hist_data[i]
57      return count
58
59  def data_load(data,filename,fieldName):
60      with open(filename) as csvfile:
61          reader = csv.DictReader(csvfile)
62          for row in reader:
63              try:
64                  data.append(float(row[fieldName]))
65              except ValueError:
66                  pass
67
68  def main():
69
70      args = sys.argv[1:]
71      filename = args[0]
72      fieldName = args[1]
73
74      bins_number = 100
75      data =[]
76
77      data_load(data,filename,fieldName)
78
79      print("\n----------\n")
80      print("%d valid values" % (len(data)))
81      print("minimum value = %.1f, maximum value = %.1f " % (min(data),max(data)))
```

```
82
83      #Equi-width histogram printing
84      equiwidth_data, equiwidth_ranges = equiwidth_histogram(data, bins_number)
85      print("equiwidth histogram:")
86      for i in range(bins_number):
87          print("[%.2f, %.2f), numtuples: %d" %(equiwidth_ranges[i],
88          equiwidth_ranges[i+1],equiwidth_data[i]))
89
90      #Equi-depth histogram printing
91      equidepth_data, equidepth_ranges = equidepth_histogram(data, bins_number)
92      print("\nequidepth histogram:")
93      for i in range(bins_number):
94          print("[%.2f, %.2f), numtuples: %d" %(equidepth_ranges[i],
95          equidepth_ranges[i+1],equidepth_data[i]))
96
97
98      #Tuples estimation
99      print("\n----------------------------------------\n")
100     print("Enter a, b ranges for estimation, press ctrl + c for exit
101      if you don't want to try multiple values.\n")
102
103     try:
104         while True:
105
106             a = int(input("Enter a: "))
107             b = int(input("Enter b: "))
108
109             if(a < b):
110                 equiwidth_est = estimate_tuples(equiwidth_data, equiwidth_ranges
111                 equidepth_est = estimate_tuples(equidepth_data, equidepth_ranges
112                 actual_tuples = len([d for d in data if a <= d < b])
113
114                 print("\nequi-width histogram estimated results: %f" % equiwidth
115                 print("equi-depth histogram estimated results: %f" % equidepth_e
116                 print("Actual results: %d\n" % actual_tuples)
117             else:
118                 print("Error! Range must be [a,b), you gave b > a. Try again.\n"
119     except KeyboardInterrupt:
120         print("\nExiting program.")
121
122 if __name__ == "__main__":
123     main()
```

2 Μέρος 1

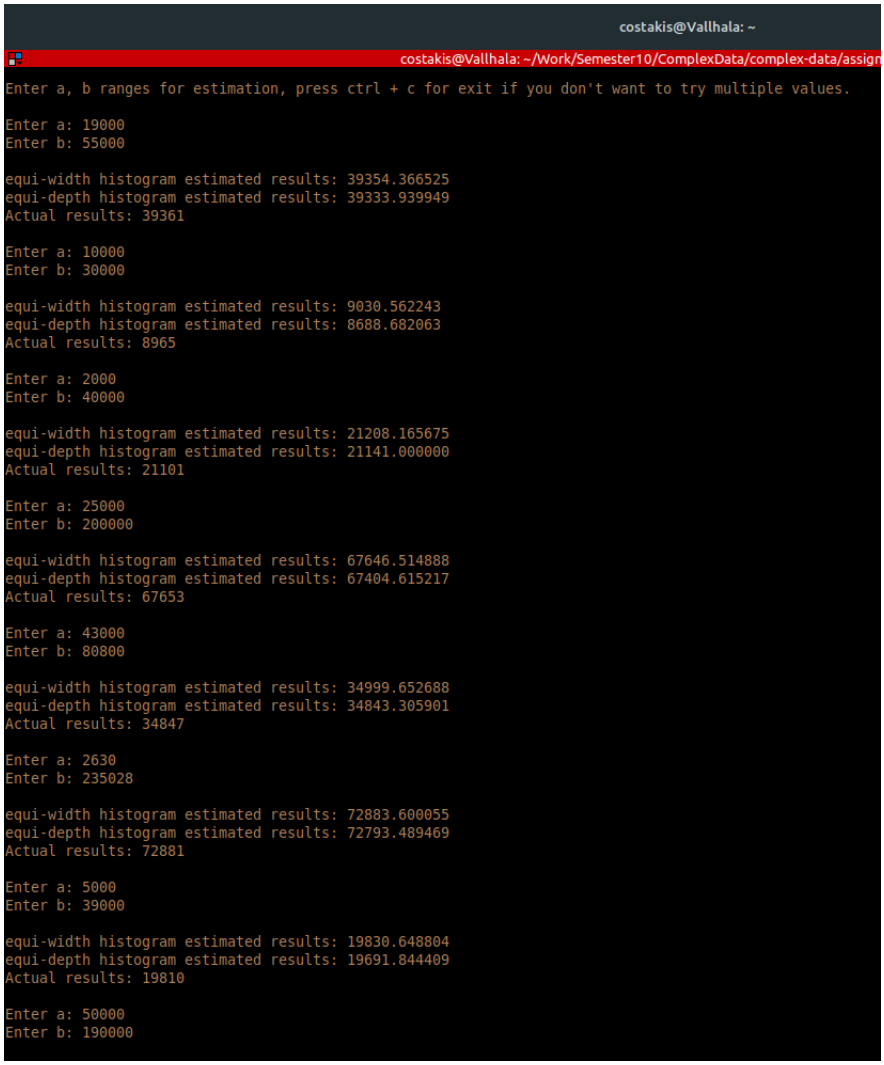
Για το πρώτο μέρος, οι συναρτήσεις που χρησιμοποιούνται για τη δημιουργία των ιστογραμμάτων equi-width & equi-depth είναι οι equiwidth_histogram() & equidepth_histogram() αντίστοιχα.
Η συνάρτηση equiwidth_histogram(data, bins) υπολογίζει το equi-width histogram ενός πίνακα δεδομένων. Αρχικά υπολογίζει την ελάχιστη και τη μέγιστη τιμή των δεδομένων και καθορίζει το πλάτος του κάθε bin με βάση τον αριθμό των bins που θα χρησιμοποιηθούν. Στη συνέχεια, δημιουργεί έναν πίνακα bin_ranges και αρχικοποιεί έναν πίνακα hist_data όπου και θα περιέχει των αριθμό των πλειάδων του κάθε range του bin. Η συνάρτηση διατρέχει τα δεδομένα και αντιστοιχίζει κάθε δεδομένο στο αντίστοιχο bin. Τέλος, επιστρέφει τα δεδομένα του ιστογράμματος και τον πίνακα των ranges του κάθε bin.
Η συνάρτηση equidepth_histogram(data, bins) υπολογίζει το equi-depth histogram ενός πίνακα δεδομένων. Αρχικά ταξινομεί τα δεδομένα και υπολογίζει το μέγεθος του bin και το υπόλοιπο με βάση τον αριθμό των bins που χρησιμοποιούνται. Στη συνέχεια, δημιουργεί έναν πίνακα bin_ranges και αρχικοποιεί έναν hist_data όπου και θα περιέχει των αριθμό των πλειάδων του κάθε range του bin. Η συνάρτηση διατρέχει τα bins και αντιστοιχίζει κάθε δεδομένο στο αντίστοιχο bin. Τέλος, επιστρέφει τα δεδομένα του ιστογράμματος και τον πίνακα των ranges του κάθε bin.

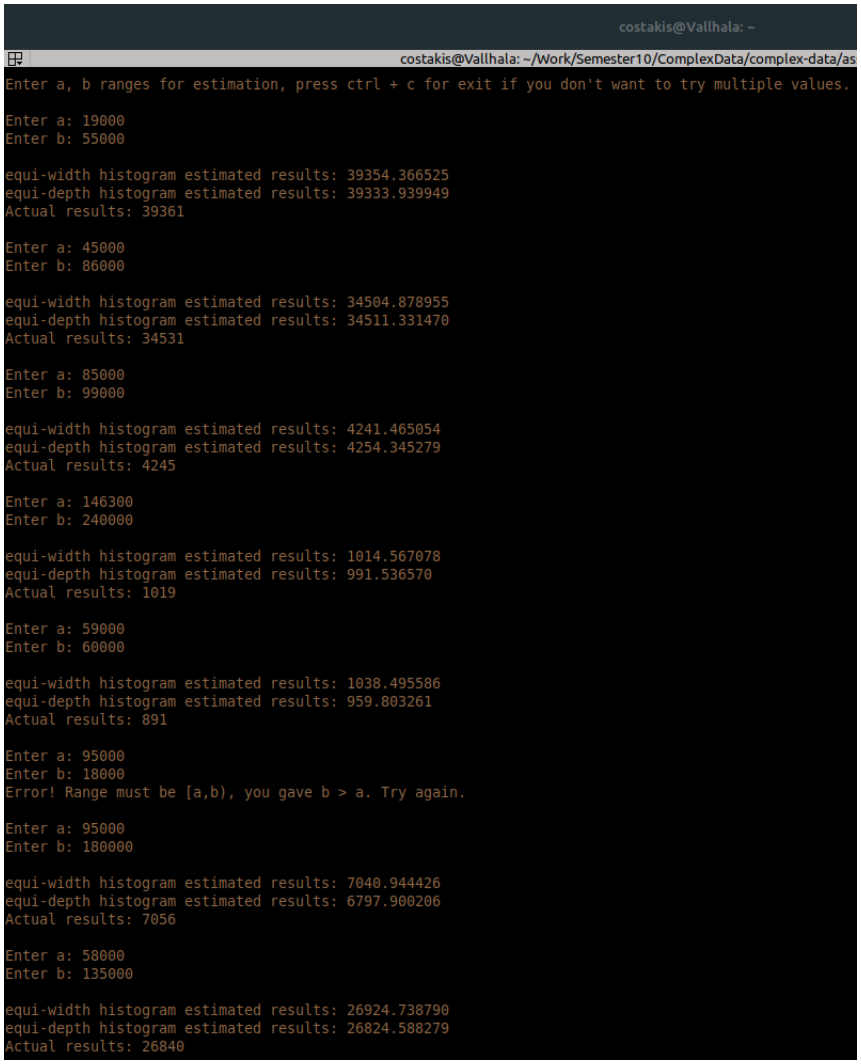
3 Μέρος 2

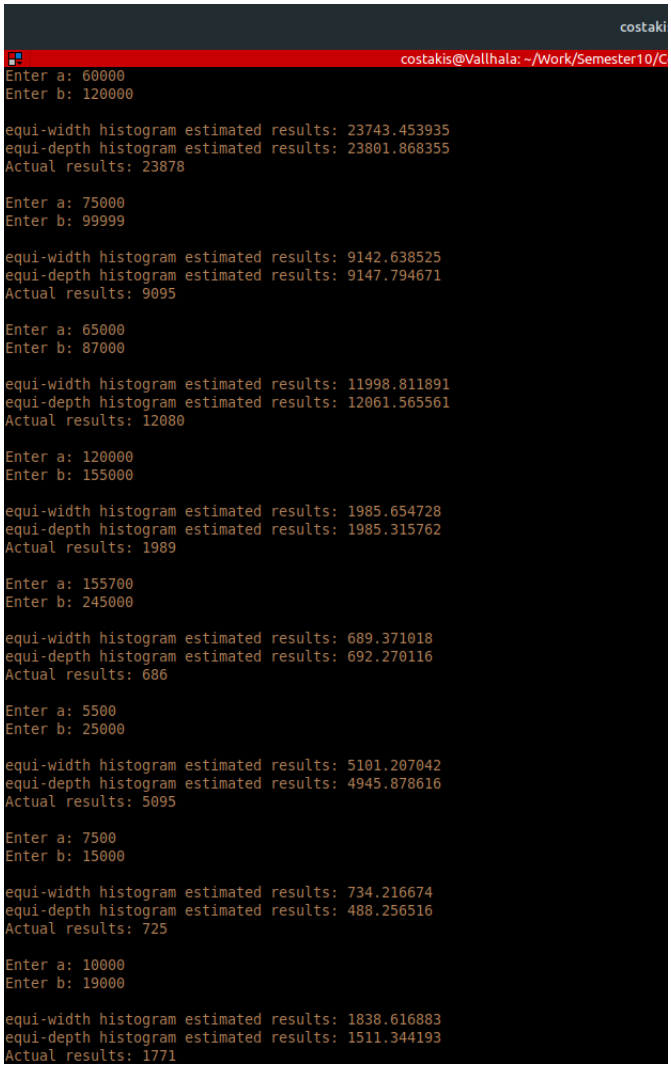
Στο δεύτερο μέρος υλοποείται επιπλέον η συνάρτηση estimate_tuples() για να εκτιμήσουμε πόσες πλειάδες έχει το αποτέλεσμα μιας ερώτησης επιλογής στο πεδίο Income, η οποία έχει σαν συνθήκη το $a \leq Income < b$.
Η συγκεκριμένη συνάρτηση υπολογίζει το πλήθος των tuples που βρίσκονται σε ένα δοσμένο εύρος [a, b), βασιζόμενη στα δεδομένα του ιστογράμματος και του πίνακα bins_ranges που είναι και ορίσματα της συνάρτησης. Αρχικά επαναλαμβάνει τον πίνακα bins_ranges για να βρει τα αντίστοιχα bins που επικαλύπτονται με το δοσμένο εύρος. Έπειτα, υπολογίζει το ποσοστό του πίνακα που εμπίπτει στο δοσμένο εύρος και το πολλαπλασιάζει με τον αριθμό των tuples σε εκείνον τον πίνακα. Τέλος, επιστρέφει το συνολικό πλήθος των tuples στο δοσμένο εύρος.

3.1 Equi-width vs Equi-depth histogram

Παρακάτω φαίνονται τα αποτελέσματα των πειραμάτων δοκιμάζοντας έναν μεγάλο αριθμό ερωτήσεων με διάφορα εύρη.




costakis@Vallhala: ~
costakis@Vallhala: ~/Work/Semester10/ComplexData/complex-data/as
Enter a, b ranges for estimation, press ctrl + c for exit if you don't want to try multiple values.
Enter a: 19000
Enter b: 55000
equi-width histogram estimated results: 39354.366525
equi-depth histogram estimated results: 39333.939949
Actual results: 39361
Enter a: 45000
Enter b: 86000
equi-width histogram estimated results: 34504.878955
equi-depth histogram estimated results: 34511.331470
Actual results: 34531
Enter a: 85000
Enter b: 99000
equi-width histogram estimated results: 4241.465054
equi-depth histogram estimated results: 4254.345279
Actual results: 4245
Enter a: 146300
Enter b: 240000
equi-width histogram estimated results: 1014.567078
equi-depth histogram estimated results: 991.536570
Actual results: 1019
Enter a: 59000
Enter b: 60000
equi-width histogram estimated results: 1038.495586
equi-depth histogram estimated results: 959.803261
Actual results: 891
Enter a: 95000
Enter b: 18000
Error! Range must be [a,b), you gave b > a. Try again.
Enter a: 95000
Enter b: 180000
equi-width histogram estimated results: 7040.944426
equi-depth histogram estimated results: 6797.900206
Actual results: 7056
Enter a: 58000
Enter b: 135000
equi-width histogram estimated results: 26924.738790
equi-depth histogram estimated results: 26824.588279
Actual results: 26840


costaki
costakis@Vallhala: ~/Work/Semester10/C
Enter a: 60000
Enter b: 120000
equi-width histogram estimated results: 23743.453935
equi-depth histogram estimated results: 23801.868355
Actual results: 23878
Enter a: 75000
Enter b: 99999
equi-width histogram estimated results: 9142.638525
equi-depth histogram estimated results: 9147.794671
Actual results: 9095
Enter a: 65000
Enter b: 87000
equi-width histogram estimated results: 11998.811891
equi-depth histogram estimated results: 12061.565561
Actual results: 12080
Enter a: 120000
Enter b: 155000
equi-width histogram estimated results: 1985.654728
equi-depth histogram estimated results: 1985.315762
Actual results: 1989
Enter a: 155700
Enter b: 245000
equi-width histogram estimated results: 689.371018
equi-depth histogram estimated results: 692.270116
Actual results: 686
Enter a: 5500
Enter b: 25000
equi-width histogram estimated results: 5101.207042
equi-depth histogram estimated results: 4945.878616
Actual results: 5095
Enter a: 7500
Enter b: 15000
equi-width histogram estimated results: 734.216674
equi-depth histogram estimated results: 488.256516
Actual results: 725
Enter a: 10000
Enter b: 19000
equi-width histogram estimated results: 1838.616883
equi-depth histogram estimated results: 1511.344193
Actual results: 1771

Από τα παραπάνω πειράματα είναι αντιληπτό πως το equi-depth histogram υπερτερεί του equi-width histogram για την εκτίμηση των πλειάδεων που έχει το αποτέλεσμα μιας ερώτησης επιλογής στο πεδίο Income, η οποία έχει σαν συνθήκη το $a \leq Income < b$. Αυτό συμβαίνει καθώς στο equi-depth histogram το κάθε bin έχει ίδιο αριθμό πλειάδων, άρα τα δεδομένα είναι όμοια μοιρασμένα στα ranges του κάθε bin. Στις εικόνες υπάρχουν και περιπτώσεις όπου το equi-width histogram δίνει καλύτερη εκτίμηση και αυτό διότι μπορεί στο συγκεκριμένο εύρος της ερώτησης, το αντίστοιχο bin του ιστογράμματος να μην επικαλύπτεται με κάποιο άλλο, με αποτέλεσμα να είναι πιο ακριβής ο αριθμός των πλειάδων.