Πανεπιστήμιο Ιωαννίνων

ΔΙΑΧΕΙΡΙΣΗ ΣΥΝΘΕΤΩΝ ΔΕΔΟΜΕΝΩΝ

ΕΡΓΑΣΙΑ 1η - Ιστογράμματα

ΝΤΟΝΤΗΣ ΒΑΣΙΛΕΙΟΣ

ΑΜ: 3300

ΜΕΡΟΣ 1ο

Τα πρώτα 15 lines,εντός του with open, είναι υπεύθυνα για το διάβασμα του csv. Κάθε γραμμή ‘’διαβάζεται’’ στο row, ως πίνακας με τα δεδομένα των στηλών, για αυτό τον λόγο στην πρώτη γραμμή που περιέχει τις κεφαλίδες του csv αποθηκεύ-ουμε το index της κεφαλίδας ‘Income’ για να διαβάζουμε μόνο την στήλη που μας ενδιαφέρει. Τα δεδομένα του Income τα αποθηκεύουμε εντός του πίνακα incomeArray. Τυπώνουμε το ελάχιστο, μέγιστο και το σύνολο των δεδομένων στον χρήστη μέσω του πίνακα incomeArray.

Equi-width: Στο equi-width histogram, όλα τα bins έχουν το ίδιο εύρος τιμών, έτσι για να βρούμε το εύρος τιμών, διαιρούμε την απόσταση του ελάχιστου προς τον μέγιστο αριθμό με τον αριθμό των bins. Αυτό το

εύρος τιμών στο πρόγραμμα είναι η μεταβλητή equiWidthBorders. Ταξινομούμε τον πίνακα με τα δεδομένα και στην συνέχεια εκτελούμε μια for για κάθε bin. Αρχικοποιούμε την μεταβλητή index ίση με 0, η οποία μας δείχνει σε ποια θέση του πίνακα βρισκόμαστε. Εντός της for, κρατάμε στην μεταβλητή equiWidthMin το αριστερό άκρο του διαστήματος του bin range, στο equiWidthMax το δεξί άκρο του ίδιου διαστήματος. Για να βρούμε το δεξί άκρο απλά προσθέτουμε το equiWidthBorders στο equiWidthMin. Μέσω της while, η οποία ελέγχει αν ένα στοιχείο του πίνακα βρίσκεται εντός του εύρος τιμών που μας ενδιαφέρει για το κάθε bin. Αν βρίσκεται εντός ανεβάζουμε κατά ένα το index για να προχωρήσουμε στον πίνακα, και ανεβάζουμε κατά 1 το numtuplesCounter που είναι η μεταβλητή που κρατάει τον αριθμό των στοιχείων που βρίσκονται εντός του bin. Όταν τελειώσουν τα στοιχεία για το συγκεκριμένο bin, τυπώνουμε στον χρήστη το εύρος τιμών και τα στοιχεία για το συγκεκριμένο bin, κάνουμε min το προηγούμενο max για να έχουμε σωστά διαστήματα στο επόμενο bin. Τέλος τυπώνουμε τα συνολικά στοιχεία που καταχωρήθηκαν σε bins, το οποίο αποτέλεσμα βγαίνει πάντα valid income values - 1, καθώς το max στοιχείο δεν καταχωρείται σε bin επειδή το τελευταίο εύρος τιμών είναι σαν το παρακάτω [x, max), έτσι δεν συμπεριλαμβάνει την max τιμή.

Equi-depth: Στο equi-depth histogram, το άθροισμα των δεδομένων που πέφτουν σε κάθε bin πρέπει να είναι το ίδιο, έτσι για να βρούμε τον αριθμό των δεδομένων που συμπεριλαμβάνεται σε κάθε bin, διαιρούμε τον αριθμό των δεδομένων με τον αριθμό των bins. Η μεταβλητή που κρατάει αυτήν την τιμή ονομάζεται numtuplesForEachBin. Εκτελούμε ξανά μια for για κάθε bin, εντός της οποίας μέσω μιας while ελέγχουμε τον αριθμό δεδομένων που θα τοποθετηθούν στο bin να μην ξεπεράσει το numtuplesForEachBin. Έχουμε ξανά τις μεταβλητές min, max για τον ορισμό του εύρους τιμών σε κάθε bin, equiDepthMin και equiDepthMax αντίστοιχα. Το max θα είναι πάντα η επόμενη θέση του πίνακα από όταν βγήκε ψευδής η συνθήκη του while. Τέλος τυπώνουμε τον αριθμό bin, το εύρος τιμών του και τον αριθμό δεδομένων που συμπεριλήφθηκαν σε αυτό, και ύστερα αλλάζουμε το min σε max, για να μπορέσει να συνεχίσει αυτοματοποιημένα ο υπολογισμός του επόμενου bin.

ΜΕΡΟΣ 2ο

Αρχικά θα αναλύσουμε τις αλλαγές στις συναρτήσεις equiwidth και equidepth από το 1ο μέρος. Και στις δυο συναρτήσεις, αρχικοποιούμε και γεμίζουμε τους πίνακες equiWidthBinRanges και equiDepthBinRanges με δυάδες που αποτυπώνουν το εύρος τιμών του bin με αριθμό (θέση στον πίνακα + 1), έτσι ώστε να γίνεται η σύγκριση των α και β με αυτά. Επίσης, αρχικοποιούμε και γεμίζουμε τους πίνακες equiWidthBinNumtuples και equiDepthBinNumtuples στα οποία αποθηκεύουμε τον αριθμό στοιχείων του κάθε bin με αριθμό (θέση στον πίνακα + 1). Εκτός αυτών, η διαδικασία είναι ίδια με το Μέρος 1ο.

Ξεκινάμε πάλι διαβάζοντας το csv και κρατάμε τα δεδομένα στον πίνακα incomeArray. Παίρνουμε τις εισόδους του χρήστη για τις μεταβλητές α και β και διατρέχοντας τον ταξινομημένο πίνακα δεδομένων βρίσκουμε πόσες τιμές βρίσκονται στο διάστημα [α, β), το σύνολο των τιμών αυτών είναι και ο αριθμός του πραγματικού αποτελέσματος που θα τυπωθεί στον χρήστη στο τέλος του προγράμματος. Αρχικοποιούμε τις μεταβλητές equiWidthEstimated και equiDepthEstimated σε 0, οι οποίες κρατάνε το εκτιμώμενο αποτέλεσμα με χρήση του equi-width histogram και equi-depth histogram αντίστοιχα. Έπειτα καλούμε τις συναρτήσεις equiwidth και equidepth για να συμπληρωθούνε με τα απαραίτητα δεδομένα οι πίνακες που θα χειριστούμε παρακάτω, και συνεχίζουμε με τον κώδικα της φωτογραφίας.

```
for i in range(len(equiWidthBinRanges)):
    if not (a > equiWidthBinRanges[i][1] or b < equiWidthBinRanges[i][0]):
        if a <= equiWidthBinRanges[i][0] and b >= equiWidthBinRanges[i][1]:
            equiWidthEstimated += equiWidthBinNumtuples[i]
        else:
            if a >= equiWidthBinRanges[i][0]:
                leftEdge = a
            else:
                leftEdge = equiWidthBinRanges[i][0]
            if b <= equiWidthBinRanges[i][1]:
                rightEdge = b
            else:
                rightEdge = equiWidthBinRanges[i][1]
            commonAreaPercent = (rightEdge - leftEdge) / (equiWidthBinRanges[i][1] - equiWidthBinRanges[i][0])
            equiWidthEstimated += commonAreaPercent * equiWidthBinNumtuples[i]
```

Διατρέχουμε το κάθε bin και ελέγχουμε τις περιπτώσεις του α και του β σε σχέση με το εύρος τιμών του bin.

1η περίπτωση: *Περίπτωση που δεν καλύπτεται ποσοστό του εύρους τιμών.*

α > δεξί μέρος διαστήματος bin. Σε αυτήν την περίπτωση δεν μας ενδιαφέρει αυτό το bin, καθώς δεν βρίσκεται εντός του ορίου που μας έδωσε ο χρήστης. (***a > equiWidthBinRanges[i][1]***, η θέση 1 αποτυπώνει το δεξιό άκρο του διαστήματος του bin.)

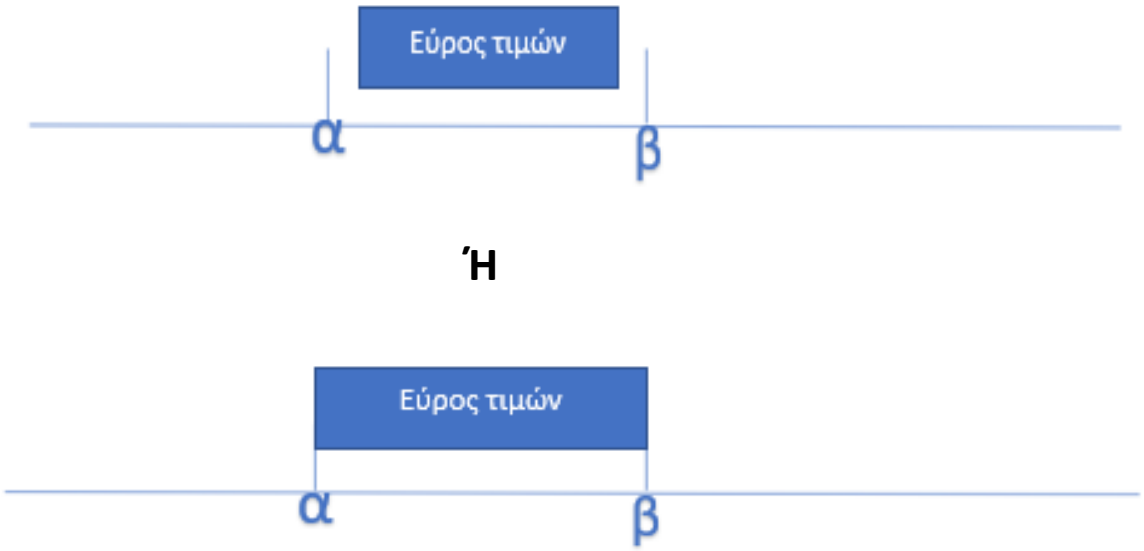
2η περίπτωση: *Περίπτωση που δεν καλύπτεται ποσοστό του εύρους τιμών.*

β < αριστερός μέρος διαστήματος bin*.** Σε αυτήν την περίπτωση δεν μας ενδιαφέρει αυτό το bin, καθώς δεν βρίσκεται εντός του ορίου που μας έδωσε ο χρήστης. (***b < equiWidthBinRanges[i][0], η θέση 0 αποτυπώνει το αριστερό άκρο του διαστήματος του bin.)

Αφού βεβαιωθούμε ότι το bin βρίσκεται εντός του [α, β), προχωράμε στις επόμενες περιπτώσεις.

3η περίπτωση: *Περιπτώσεις που καλύπτεται το 100% του εύρους τιμών.*

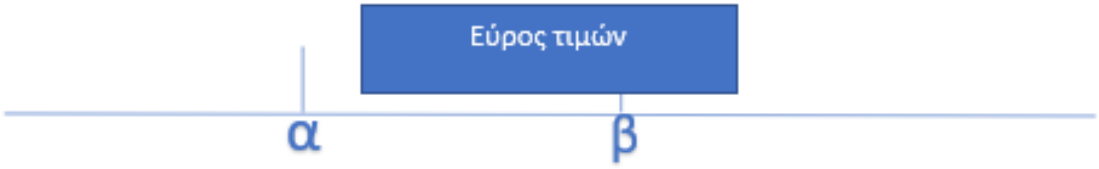
α <= αριστερός μέρος διαστήματος bin και β >= δεξιό μέρος διαστήματος bin. Σε αυτήν την περίπτωση μας ενδιαφέρει αυτό το bin, και αφού το διάστημα [<=α, >=β) καλύπτει σίγουρα το 100% του διαστήματος [α, β), στο estimated αποτέλεσμα προσθέτουμε τον αριθμό των numtuples αυτού του bin το οποίο είναι αποθηκευμένο στον πίνακα equiWidthBinNumtuples[i]. (***α <= equiWidthBinRanges[i][0] και b >= equiWidthBinRanges[i][1]***)



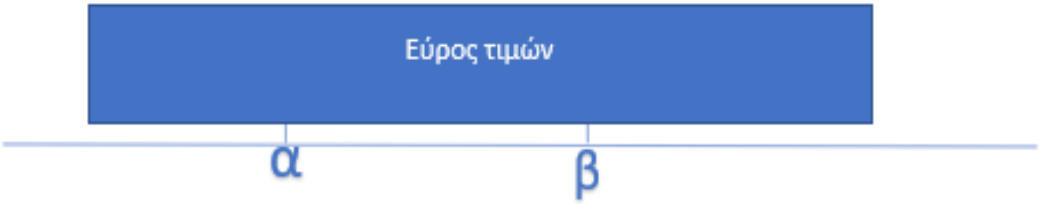
4η περίπτωση: *Περιπτώσεις που καλύπτεται ένα ποσοστό του εύρους τιμών.*

Για να μπορέσουμε να υπολογίσουμε το ποσοστό της κοινής περιοχής τιμών μεταξύ του εύρους τιμών και του διαστήματος [α, β), χρειάζεται να βρούμε τον μεγαλύτερο αριθμό εκ των α, αριστερό άκρο διαστήματος εύρους ζώνης και τον μικρότερο αριθμό εκ των β, δεξιό άκρο διαστήματος εύρους ζώνης. Αφού βρούμε τον αριθμό κοινού διαστήματος, τον διαιρούμε με το πλήθος του εύρους τιμών, και έτσι βρίσκουμε το ποσοστό κοινής περιοχής. Αφού έχουμε το ποσοστό, το πολλαπλασιάζουμε με τον αριθμό πλειάδων για να βρούμε το επιθυμητό εκτιμώμενο αριθμό δεδομένων.

Ή



Ή



Ακριβώς την ίδια δουλειά κάνει και η επόμενη for, απλά με τις τιμές του equi-depth.

Τέλος, θα δείξουμε κάποια αποτελέσματα πειραμάτων διάφορων εισόδων, για να καταλήξουμε στην σύγκριση μεταξύ των δύο ιστογραμμάτων.

ΠΕΙΡΑΜΑΤΑ

Πείραμα 1ο: ***α:19000 , β:55000***

```
Please give the number of a: 19000
Please give the number of b: 55000

Equiwidth estimated results: 39354.36652460602
Equidepth estimated results: 39333.939948818304
Actual results: 39361
```

Αποτελέσματα 1ου Πειράματος:

Equiwidth estimated results: ***39354.36652460602***

Equidepth estimated results: ***39333.939948818304***

Actual results: ***39361***

Καλύτερος υπολογισμός: Equi-Width

Πείραμα 2ο: ***α:5000 , β: 10000***

```
Please give the number of a: 5000
Please give the number of b: 10000

Equiwidth estimated results: 108.853286151321
Equidepth estimated results: 298.6970417110547
Actual results: 109
```

Αποτελέσματα 2ου Πειράματος:

Equiwidth estimated results: ***108.853286151321***

Equidepth estimated results: ***298.6970417110547***

Actual results: ***109***

Καλύτερος υπολογισμός: Equi-Width

Πείραμα 3ο: ***α:2500 , β:100000***

```
Please give the number of a: 2500
Please give the number of b: 100000

Equiwidth estimated results: 66863.03389548181
Equidepth estimated results: 66873.09375
Actual results: 66804
```

Αποτελέσματα 3ου Πειράματος:

Equiwidth estimated results: 66863.03389548181

Equidepth estimated results: 66873.09375

Actual results: 66804

Καλύτερος υπολογισμός: Equi-Width

Πείραμα 4ο: ***α:100000 , β:250000***

```
Please give the number of a: 100000
Please give the number of b: 250000

Equiwidth estimated results: 6036.96610451818
Equidepth estimated results: 6026.90625
Actual results: 6097
```

Αποτελέσματα 4ου Πειράματος:

Equiwidth estimated results: 6036.96610451818

Equidepth estimated results: 6026.90625

Actual results: 6097

Καλύτερος υπολογισμός: Equi-Width

Πείραμα 5ο: ***α:100000 , β:125000***

```
Please give the number of a: 100000
Please give the number of b: 125000

Equiwidth estimated results: 3848.6956882086965
Equidepth estimated results: 3824.049585192415
Actual results: 3894
```

Αποτελέσματα 5ου Πειράματος:

Equiwidth estimated results: 3848.6956882086965

Equidepth estimated results: 3824.049585192415

Actual results: 3894

Καλύτερος υπολογισμός: Equi-Width

Πείραμα 6ο: ***α:25000 , β:25010***

```
Please give the number of a: 25000
Please give the number of b: 25010

Equiwidth estimated results: 7.87359987649254
Equidepth estimated results: 7.029893924783028
Actual results: 25
```

Αποτελέσματα 6ου Πειράματος:

Equiwidth estimated results: 7.87359987649254

Equidepth estimated results: 7.029893924783028

Actual results: 25

Καλύτερος υπολογισμός: Equi-Width

Πείραμα 7ο: ***α:10000 , β:200000***

```
Please give the number of a: 10000
Please give the number of b: 200000

Equiwidth estimated results: 72643.24669800398
Equidepth estimated results: 72081.18419282603
Actual results: 72642
```

Αποτελέσματα 7ου Πειράματος:

Equiwidth estimated results: 72643.24669800398

Equidepth estimated results: 72081.18419282603

Actual results: 72642

Καλύτερος υπολογισμός: Equi-Width

ΣΥΜΠΕΡΑΣΜΑ

Σε όλα τα πειράματα καλύτερο αποτέλεσμα είχε η χρήση του equi-width histogram και σε μεγάλα αλλά και σε μικρά εύρη τιμών. Άρα καταλήγουμε στο ότι το equi-width histogram υπερτερεί του equi-depth histogram.