ΜΑΘΗΜΑ: "Εισαγωγή στον Προγραμματισμό"

ΕΡΓΑΣΤΗΡΙΟ 6: Περιγραφική Στατιστική

Άσκηση 1. Σε προηγούμενο εργαστήριο φτιάξατε ένα data frame, το οποίο και είχατε αποθηκεύσει ως dataLab5.txt. Το αρχείο αυτό υπάρχει στο eclass του μαθήματος. Ανοίξτε την R και φορτώστε το (με κατάλληλο τρόπο). **Υπόδειξη:** Χρησιμοποιήστε την εντολή read.table και εισάγετε τα δεδομένα με τον τίτλο datafile 1. Στη συνέχεια με την εντολή attach φορτώστε τα δεδομένα στην R ώστε να είναι έτοιμα για επεξεργασία

(i). Από τις μεταβλητές στο πλαίσιο δεδομένων, ποιες είναι ποσοτικές και ποιες ποιοτικές;

(ii). Να συμπληρώσετε τον παρακάτω πίνακα

Στατιστικό Περιγραφικό Μέτρο	**Χοληστερίνη**	**Βάρος**	**Πίεση**
Μέση τιμή			
Τυπική απόκλιση			
1ο τεταρτημόριο			
Διάμεσος			
3ο τεταρτημόριο			
Εύρος			
Ενδοτεταρτημοριακό εύρος			
Περικομμένος μέσος 10%			
80% ποσοστιαίο σημείο			
35% ποσοστιαίο σημείο			
Συντελεστής γραμμικής συσχέτισης			

(iii). Εξηγήστε το αποτέλεσμα που σας δίνει η εντολή

> summary(datafile1)

(iv). Τι ποσοστό των ατόμων του δείγματος, έχει βάρος μικρότερο από 70kg; Τι ποσοστό έχει ύψος μεγαλύτερο από 175cm; Σε κάθε μια από τις παραπάνω υποομάδες, πόσοι είναι άνδρες, πόσοι γυναίκες;

Υπόδειξη: Να χρησιμοποιήσετε τις εντολές which και length.

Άσκηση 2 (Άσκηση 10, Ντζούφρας & Καρλής). Δημιουργήστε τα δεδομένα του παρακάτω Πίνακα στην R. Στη συνέχεια:

(i) Υπολογίστε τη μέση τιμή για κάθε μάθημα ως προς το φύλο.

(ii) Βρείτε τη μέγιστη βαθμολογία για κάθε μάθημα ξεχωριστά.

(iii) Βρείτε τη μέγιστη βαθμολογία που πήρε φοιτητής σε οποιοδήποτε μάθημα.

(iv) Βρείτε το μέσο όρο της βαθμολογίας των φοιτητών και την κατάταξή τους, με βάση αυτό το μέσο όρο στο σύνολο των φοιτητών, αλλά και στο σύνολο των φοιτητών από το ίδιο έτος.

(v) Ποιος είναι ο καλύτερος φοιτητής; Τυποποιήστε τους βαθμούς ώστε να είναι συγκρίσιμοι και υπολογίστε το μέσο όρο με τη χρήση των τυποποιημένων βαθμών.

(vi) Ποιο ποσοστό φοιτητών πέρασε όλα τα μαθήματα;

(vii) Ποιος είναι ο μέσος βαθμός και η διακύμανση για αυτούς που πέρασαν κάθε μάθημα;

(viii) Κατασκευάστε από τα δεδομένα έναν πίνακα. Πώς θα υπολογίσετε τώρα τη μέση τιμή κάθε μαθήματος ως προς το φύλο αλλά και τη μέση τιμή για κάθε μάθημα ξεχωριστά;

(ix) Για κάθε μάθημα υπολογίστε το λόγο της τυπικής απόκλισης με τη μέση τιμή. Τι μας δείχνει αυτός ο λόγος;

Χημεία	Φυσική	Μαθηματικά	Αρχαία	Φύλο	Έτος
93	42	98	34	Α	1
71	67	68	33	Α	1
77	59	36	24	Α	1
78	70	92	24	Α	1
77	59	44	31	Α	1
81	50	45	22	Α	2
88	50	58	23	Γ	2
74	51	31	32	Γ	2
67	45	70	31	Γ	2
78	64	46	26	Γ	2
77	49	41	75	Α	1
67	49	46	81	Α	1
63	48	65	87	Γ	1
83	51	62	100	Γ	1
73	56	20	81	Γ	1
70	47	22	100	Γ	2
78	53	92	77	Α	2
95	56	56	89	Α	2
88	49	28	100	Α	2
75	71	94	77	Α	2

Άσκηση 3 (Φουσκάκης 2013, 2.14) Με τη βοήθεια της εντολής scan(), να διαβάσετε τα δεδομένα του λινκ http://www.math.ntua.gr/~fouskakis/Rbook/data1-da.txt και να τα αποθηκεύσετε σε έναν πίνακα Χ με 6 στήλες.

(i) Συγκρίνεται τα στοιχεία του πίνακα Χ με τα αρχικά δεδομένα και βρείτε τη διάσταση του πίνακα που δημιουργήσατε.

(ii) Αποθηκεύστε τη δεύτερη και Τρίτη μόνο στήλη του πίνακα σε ένα αρχείο .txt (έστω HW0604.txt) στο σκληρό σας δίσκο. Επιβεβαιώστε ότι ξέρετε που αποθηκεύτηκε.

(iii) Μετατρέψτε τον πίνακα X σε πλαίσιο δεδομένων και δώστε τα εξής ονόματα στις στήλες του: ID, AGE, FEV, HEIGHT, SEX, SMOKING.

(iv) Βρείτε το είδος των μεταβλητών HEIGHT και SMOKING του πλαισίου δεδομένων κάνοντας χρήση της εντολής mode() (δείτε πληροφορίες για την εντολή help(mode)). Στη συνέχεια, δηλώστε στην R ότι οι μεταβλητές SEX και SMOKING είναι κατηγορικές. Δείτε ποιες είναι οι κατηγορίες τους.

(v) Αλλάξτε τις τιμές της μεταβλητές SMOKING ως ακολούθως: 1=ΝΑΙ, 0=ΟΧΙ. Με τη βοήθεια της εντολής subset(), επιλέξτε μόνο τις γραμμές του πλαισίου που αντιστοιχούν σε καπνιστές.

Άσκηση 4 (Φουσκάκης 2013, άσκηση 3.3): Στη σελίδα http://www.math.ntua.gr/~fouskakis/Rbook/plants.txt βρίσκονται δεδομένα από 28 περιοχές της Ελλάδας στις οποίες ευδοκιμεί ένα φυτό συγκεκριμένου είδους. Η 1η στήλη περιλαμβάνει το ύψος του (σε μέτρα) και η 2η τη γεωγραφική του θέση (1: Β. Ελλάδα, 2: Ν. Ελλάδα). Αφού δείτε τα δεδομένα του αρχείου, να τα περάσετε στην R απευθείας με χρήση του link που σας δίνεται. Να περαστούν ως data frame με όνομα df1 και να δώσετε τα ονόματα height, area στις δύο μεταβλητές. Η 2η να είναι παράγοντας.

(i) Να υπολογιστούν ο δειγματικός μέσος, η δειγματική διάμεσος, η δειγματική τυπική απόκλιση καθώς και το 1ο και το 3ο τεταρτημόριο του ύψους των φυτών. Επίσης να υπολογιστεί η δειγματική διασπορά του ύψους των φυτών.

(ii) Να κατασκευαστεί το ιστόγραμμα συχνοτήτων και σχετικών συχνοτήτων του ύψους των φυτών. Να πειραματιστείτε με το πλήθος των κλάσεων. Να βρεθούν οι κλάσεις σύμφωνα με τον κανόνα Sturges αλλά και σύμφωνα με τον κανόνα των Freedman-Diaconis.

(iii) Να κατασκευαστούν στο ίδιο γράφημα δύο θηκογράμματα που να αφορούν το ύψος των φυτών στη Β. Ελλάδα και στη Ν. Ελλάδα, αντίστοιχα. Δώστε κατάλληλα ονόματα στα θηκογράμματα. Τι παρατηρείτε;

(iv) Να υπολογιστεί η διάμεσος του ύψους των φυτών ξεχωριστά για τη Β. Ελλάδα και για τη Ν. Ελλάδα, με δύο τρόπους. Ο 2ος τρόπος μπορεί να στηρίζεται στην εντολή by() (πληκτρολογήστε helo(by) για περισσότερες πληροφορίες). Τι παρατηρείτε σε συνδυασμό και με το αποτέλεσμα του (iii);