Ε Θ Ν Ι Κ Ο Μ Ε Τ Σ Ο Β Ι Ο Π Ο Λ Υ Τ Ε Χ Ν Ε Ι Ο
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ & ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
Τομέας Μαθηματικών
Πολυτεχνειούπολη – Ζωγράφου ΑΘΗΝΑ - 157 80
ΤΗΛ. : 772 1774
FAX : 772 1775

ΜΑΘΗΜΑ: *Ανάλυση Δεδομένων με Η/Υ (6ο εξάμηνο)*
ΔΙΔΑΣΚΩΝ: *Δημήτρης Φουσκάκης*

ΕΡΓΑΣΙΑ 1η

Θέμα Εργασίας: Εισαγωγή στην R και Περιγραφική Στατιστική

Άσκηση 1

Αλυσίδα φαρμακείων θέλει να μελετήσει τους παράγοντες που επηρεάζουν το ποσό χρημάτων που θα ξοδέψει ένας πελάτης της σε αγορά κάποιου μη φαρμακευτικού προϊόντος της. Για αυτό το λόγο συλλέχθηκαν πληροφορίες από 72 (διαφορετικούς) πελάτες της που έκαναν μία μόνο αγορά τον προηγούμενο μήνα για κάποιο μη φαρμακευτικό προϊόν. Οι μεταβλητές που συλλέχθηκαν στο δείγμα για κάθε πελάτη είναι οι εξής: (α) η ηλικία (**age**) σε έτη, (β) η κατηγορία **(category)** του μη φαρμακευτικού προϊόντος αγοράς, με κλάσεις τις "*healthcare*" (υγειονομικά), "*cosmetics*" (καλλυντικά) και "*other*" (άλλο), (γ) το φύλο (**sex)**, με κλάσεις τις "*Man*" (Άνδρας) και "*Woman*" (Γυναίκα), (δ) το ποσό χρημάτων **(med)**, σε ευρώ, που ξόδεψε στο φαρμακείο για φαρμακευτικά προϊόντα τον προηγούμενο μήνα, (ε) τον αριθμό κατοίκων **(population)** που έχει η περιοχή που βρίσκεται το κατάστημα που έγινε η αγορά και (στ) το ποσό χρημάτων (**money**) σε ευρώ που ξόδεψε στην αγορά του μη φαρμακευτικού προϊόντος.

Τα δεδομένα βρίσκονται στο αρχείο:

http://www.math.ntua.gr/~fouskakis/Data_Analysis/Exercises/pharmacy.txt

Στα δεδομένα μας, με τον χαρακτήρα "$" συμβολίζουμε τις αγνοούμενες τιμές. Εισάγετε τα δεδομένα στην R με χρήση της εντολής read.table, αλλάζοντας κατάλληλα το σύμβολο για τις αγνοούμενες τιμές (για αυτόματη αλλαγή, ελέγξτε στο μενού help της R τα δυνατά ορίσματα της εντολής read.table). Ποια είναι η δομή του

αντικειμένου που δημιουργείται από την παραπάνω εντολή; Στη συνέχεια αφαιρέστε οποιαδήποτε γραμμή περιέχει αγνοούμενη τιμή.

i. Δώστε μια περιγραφική ανάλυση, για τις τιμές ή κλάσεις (κατηγορίες) κάθε μίας από τις 6 μεταβλητές ξεχωριστά, η οποία να αποτελείται από κατάλληλες αριθμητικές και γραφικές μεθόδους και σχολιάστε τα ευρήματα σας.
ii. Με τη βοήθεια κατάλληλου διαγράμματος εξετάστε περιγραφικά αν το ποσό (**money**) που ξοδεύει κάποιος πελάτης σε μη φαρμακευτικά προϊόντα στο δείγμα διαφοροποιείται ανάλογα με την κατηγορία αγορών (**category**) σε μη φαρμακευτικά προϊόντα. Υλοποιήστε παρόμοιες συγκρίσεις, με χρήση διαγραμμάτων, μεταξύ των τιμών της μεταβλητής **money** και των τιμών ή κλάσεων καθεμιάς από τις υπόλοιπες μεταβλητές. Τι συμπεραίνετε;
iii. Να κατασκευαστεί ο πίνακας συχνοτήτων και σχετικών συχνοτήτων για τα δεδομένα που αφορούν την ηλικία των ατόμων που συμμετέχουν στην έρευνα με τη χρήση 3 κλάσεων: [18-30), [30-50), [50 και άνω). Δώστε κατάλληλα ονόματα στις κατηγορίες της νέας αυτής μεταβλητής την οποία ονομάστε την **f_age**. Εν συνεχεία, κατασκευάστε μια ακόμα κατηγορική μεταβλητή, με όνομα **f_pop**, με κλάσεις $[0,q_1)$, $[q_1,q_2)$, $[q_2,q_3)$, [q_3 και άνω), όπου q_i (i = 1, 2, 3) είναι το i-στο τεταρτημόριο των τιμών της μεταβλητής **population**. Κατασκευάστε έναν δισδιάστατο πίνακα συχνοτήτων των μεταβλητών **f_pop** και **f_age** στο δείγμα. Δώστε τις σχετικές συχνότητες κελιών και σχολιάστε τα αποτελέσματα. Δημιουργείστε ένα στοιβαγμένο ραβδόγραμμα και σχολιάστε τα αποτελέσματα.

Άσκηση 2

α) Τέλειος αριθμός ονομάζεται ένας φυσικός αριθμός του οποίου το άθροισμα των διαιρετών του, εκτός του εαυτού του, είναι ίσο με τον αριθμό αυτόν (π.χ. οι διαιρέτες του 6 είναι οι 1, 2, 3 και το άθροισμα αυτών είναι ίσο με 6). Ένας τρόπος εύρεσης των διαιρετών ενός θετικού ακεραίου αριθμού n, είναι να ελέγξετε με ποιους αριθμούς (πέραν του εαυτού του) i = 1,…,n-1 το υπόλοιπο της διαίρεσης του είναι 0. Με χρήση των προηγούμενων πληροφοριών να γραφτεί συνάρτηση στην R, με όνομα της επιλογή σας, που θα δέχεται ένα διάνυσμα από θετικούς ακέραιους αριθμούς και θα εξάγει ένα διάνυσμα που θα περιέχει μόνο όσους εξ αυτών είναι τέλειοι. Σε περίπτωση που κανείς από τους αριθμούς του αρχικού διανύσματος δεν είναι τέλειος θα εκτυπώνει στην οθόνη του υπολογιστή σχετικό μήνυμα. Η συνάρτηση αρχικά θα

ελέγχει αν το αρχικό διάνυσμα περιέχει μόνο θετικούς ακέραιους αριθμούς, σε διαφορετική περίπτωση θα εμφανίζει μήνυμα σφάλματος και θα τερματίζει. **Παρατήρηση**: Ο αριθμός 1 δεν είναι τέλειος αριθμός.

β) Να γραφτεί συνάρτηση στην R με όνομα της επιλογή σας, η οποία θα δέχεται ως όρισμα θετικό ακέραιο αριθμό n και θα προσομοιώνει n τυχαίες τιμές από την Εκθετική κατανομή με μέση τιμή 1/2. Η προσομοίωση κάθε τιμής θα γίνεται με χρήση της παρακάτω μεθοδολογίας: Αρχικά προσομοιώνουμε μια τυχαία τιμή $u \sim U(0,1)$ (με χρήση της εντολής runif) και στη συνέχεια επιλύουμε τη σχέση $F(x) = u$ ως προς x, όπου με $F(x)$ συμβολίζουμε τη συνάρτηση κατανομής της Εκθετικής κατανομής με μέση τιμή 1/2. Η τιμή x που προκύπτει είναι μια προσομοιωμένη τυχαία τιμή από την εν λόγω Εκθετική κατανομή. Η συνάρτηση αρχικά θα ελέγχει αν ο αριθμός n είναι θετικός ακέραιος, διαφορετικά θα εκτυπώνει κατάλληλο μήνυμα λάθους και θα τερματίζει.

γ) Με χρήση της συνάρτησης του β) ερωτήματος προσομοιώστε 4 διαφορετικά δείγματα, μεγέθους $n_1 = 10$, $n_2 = 100$, $n_3 = 1000$ και $n_4 = 10000$, από την Εκθετική κατανομή με μέση τιμή 1/2 και στη συνέχεια για καθένα από αυτά τα δείγματα να κάνετε το αντίστοιχο ιστόγραμμα (συνολικά 4), στο οποίο θα εμφανίζεται και το γράφημα της συνάρτησης πυκνότητας πιθανότητας της αντίστοιχης Εκθετικής κατανομής.

Οδηγίες

- Η εργασία θα πρέπει **να παραδοθεί ηλεκτρονικά** στον **Γιώργο Τζουμέρκα** στο email του, tzoum_giorgos@hotmail.gr.
- Η εργασία που θα παραδώσετε πρέπει να είναι **σε pdf μορφή**. Παρακαλώ χρησιμοποιήστε τον **ακόλουθο τίτλο στο pdf αρχείο σας**: Surname-Name-Ex1.pdf, όπου Surname είναι το επώνυμό σας (με λατινικούς χαρακτήρες) και Name το όνομα σας (με λατινικούς χαρακτήρες). Π.χ. αν παρέδιδα εγώ εργασία θα την ονόμαζα ως εξής: Fouskakis-Dimitris-Ex1.pdf.
- Παρακαλώ χρησιμοποιήστε **ένα εξώφυλλο στο pdf αρχείο σας**, στο οποίο να αναγράφεται ο τίτλος της εργασίας (Εισαγωγή στην R και Περιγραφική Στατιστική), **το ονοματεπώνυμό σας, το email σας, η Σχολή σας και ο αριθμός μητρώου σας**.

- Θα πρέπει να **αποστείλετε ένα μόνο αρχείο**. Η εργασία θα πρέπει να περιλαμβάνει τους κώδικες της R, όχι σε παράρτημα αλλά στην απάντηση του κάθε ερωτήματος.
- Η εργασία θα πρέπει να αποσταλεί **μέχρι την Παρασκευή 29/03/2024 στις 13:00. Καμιά εργασία δεν θα γίνει δεκτή μετά την ώρα αυτή.**
- Η εργασία θα πρέπει να είναι σε μορφή **αναφοράς** και να περιλαμβάνει τους κώδικες της R με πλήρη επεξήγηση, γραφήματα και πίνακες με κατάλληλους τίτλους και πλήρη επεξήγηση των αποτελεσμάτων. Επίσης, δε θα πρέπει να υπερβαίνει τις **15 σελίδες** με μέγεθος γραμματοσειράς **12**.
- Θα δοθεί ιδιαίτερη σημασία **στην παρουσίαση της εργασίας**. Η εργασία πρέπει να είναι κατανοητή και να περιγράφει οτιδήποτε χρησιμοποιήσατε πειστικά για κάποιον που δεν γνωρίζει πολλά για το αντικείμενο.

Εύχομαι Επιτυχία