

**ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ****ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ & ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ****Τομέας Μαθηματικών****Πολυτεχνειούπολη – Ζωγράφου ΑΘΗΝΑ - 157 80****ΤΗΛ. : 772 1774****FAX : 772 1775****ΜΑΘΗΜΑ:** *Ανάλυση Δεδομένων με H/Y (6^ο εξάμηνο)***ΔΙΔΑΣΚΩΝ:** *Δημήτρης Φουσκάκης*

ΕΡΓΑΣΙΑ 2^η

Θέμα Εργασίας: Στατιστική Συμπερασματολογία

Άσκηση 1

Διαφημιστική εταιρεία μελετά τους παράγοντες που επηρεάζουν το ύψος των χρημάτων που χρησιμοποιεί για αγορές ένας ιδιώτης κατά τις εορταστικές περιόδους. Για αυτό το λόγο, συλλέχθηκαν πληροφορίες σε δείγμα 53 διαφορετικών ανθρώπων για τις ακόλουθες μεταβλητές: την ηλικία (**age**) του ατόμου σε έτη, την μοναδική εορταστική περίοδο (**holiday**) που πραγματοποιήθηκαν οι αγορές, με κατηγορίες τις "Christmas", "Easter", "Other", (Χριστούγεννα – Πάσχα – Λοιπές), το φύλο (**sex**) του ατόμου που έκανε τις αγορές, με κατηγορίες τις "Man", "Woman", (Αντρας – Γυναίκα), το μέσο χρόνο (**time**) σε λεπτά που θεωρεί πως γίνεται αποδέκτης διαφημιστικών προωθήσεων (από οποιοδήποτε μέσο) στη διάρκεια μιας ημέρας, το μηνιαίο του εισόδημα (**salary**) σε ευρώ και τέλος το ποσό (**spend**) σε ευρώ που ξόδεψε την εορταστική περίοδο. Κάθε άνθρωπος στο δείγμα έχει προβεί αναγκαστικά σε αγορές σε μία μόνο περίοδο από τις τρεις της μεταβλητής **holiday**.

Τα δεδομένα βρίσκονται στο αρχείο:

http://www.math.ntua.gr/~fouskakis/Data_Analysis/Exercises/gifts.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. Επιπλέον αφαιρέστε, αν υπάρχει, οποιαδήποτε γραμμή περιέχει τιμές από μη ενήλικα άτομα (ηλικία μικρότερη των 18 ετών).

- i. Ο προϊστάμενος του τμήματος που έχει αναλάβει τη στατιστική έρευνα υποστηρίζει ότι οι καταναλωτές, κατά τις εορταστικές περιόδους ξοδεύουν, κατά μέσο όρο, περισσότερα από 500 ευρώ. Αφού πρώτα οριστεί κατάλληλος έλεγχος, (όσον αφορά την τ.μ. **spend**), να ελεγχθεί σε ε.σ. 1% ο συγκεκριμένος έλεγχος υπόθεσης. Να αιτιολογήσετε όποια μέθοδο χρησιμοποιηθεί ως κατάλληλη.
- ii. Στην ομάδα που διενεργεί την έρευνα υπάρχει η πεποίθηση ότι τα Χριστούγεννα οι καταναλωτές φαίνεται να ξοδεύουν περισσότερα σε σύγκριση με τις άλλες εορταστικές περιόδους. Να κατασκευάσετε μια νέα κατηγορική μεταβλητή (**Christmas**) με δύο κατηγορίες (0: εορταστική περίοδος εκτός Χριστουγέννων και 1: εορταστική περίοδος Χριστουγέννων) και περιγράψτε κατάλληλα τις τιμές της τ.μ. **spend** ξεχωριστά για κάθε κατηγορία της νέας αυτής μεταβλητής (συμπεριλάβετε γραφικές και αριθμητικές μεθόδους). Σχολιάστε τα ευρήματά σας.
- iii. Να ελεγχθεί σε ε.σ. 5% ο ισχυρισμός της ομάδας που διενεργεί την έρευνα που αναφέρθηκε στο ερώτημα ii. Να αιτιολογήσετε όποια μέθοδο χρησιμοποιηθεί ως κατάλληλη.
- iv. Ο προϊστάμενος του τμήματος που έχει αναλάβει τη στατιστική έρευνα υποστηρίζει ότι δεν υπάρχει διαφορά στο μέσο ποσό χρημάτων που ξοδεύονται κατά την περίοδο των εορτών (**spend**), σε σχέση με το φύλο του ατόμου (**sex**). Να οριστεί κατάλληλος έλεγχος υπόθεσης και έπειτα να απαντηθεί το ερώτημα αυτό. Επιλέξτε ε.σ. 10%.
- v. Να κατασκευάσετε μια νέα κατηγορική μεταβλητή (**spend_bin**) με δύο κατηγορίες (0: μικρό ποσό και 1: μεγάλο ποσό) η οποία να είναι 0 όταν **spend** < 673 ευρώ και 1 διαφορετικά. Έστω ότι θέλετε να ελέγξετε αν η πιθανότητα να ξοδευτεί μεγάλο ποσό την περίοδο των εορτών είναι μεγαλύτερη στις γυναίκες σε σχέση με τους άνδρες. Να οριστεί κατάλληλος έλεγχος υπόθεσης και έπειτα να απαντηθεί το ερώτημα αυτό. Επιλέξτε ε.σ. 2%. Να αιτιολογήσετε όποια μέθοδο χρησιμοποιηθεί ως κατάλληλη για να απαντήσετε το ερευνητικό αυτό ερώτημα.

- vi. Να ελεγχθεί σε ε.σ. 5% η υπόθεση ότι το φύλο του ατόμου (**sex**) είναι ανεξάρτητο της εορταστικής περιόδου που έγιναν οι αγορές (**holiday**).

Άσκηση 2

- 1) Θεωρήστε ότι οι παρακάτω παρατηρήσεις μας δίνουν τον αριθμό προσπαθειών 10 αθλητών του τμήματος μπάσκετ ενός δημοτικού σχολείου μέχρι να επιτύχουν πρώτη φορά τρίποντο από μια συγκεκριμένη θέση.

3 2 1 5 2 4 5 2 3 10

- Να βρεθεί αναλυτικά η Ε.Μ.Π. του ποσοστού επιτυχίας της ομάδας στα τρίποντα από αυτή τη θέση.
 - Να βρεθεί γραφικά η Ε.Μ.Π. του ποσοστού επιτυχίας της ομάδας στα τρίποντα από αυτή τη θέση.
- 2) Να γραφτεί συνάρτηση στην R η οποία να εκτιμά την πιθανότητα σε τρεις ρίψεις ενός (τίμιου-εξάεδρου) ζαριού, η πρώτη ρίψη και η τρίτη να είναι ο ίδιος αριθμός, ενώ ο αριθμός της δεύτερης ρίψης να είναι διαφορετικός (από τις άλλες δύο). Πιο συγκεκριμένα, η συνάρτηση θα επαναλαμβάνει την διαδικασία ρίψης των τριών ρίψεων του ζαριού, N φορές και στο τέλος θα εξάγει τη σχετική συχνότητα των φορών που συνέβη το επιθυμητό αποτέλεσμα. Ως είσοδο η συνάρτηση θα δέχεται τον θετικό ακέραιο αριθμό N (σε περίπτωση που το N δεν είναι θετικός ακέραιος η συνάρτηση θα επιστρέφει μήνυμα λάθους και θα τερματίζει). Για την υλοποίηση της προσομοίωσης μπορείτε να χρησιμοποιήσετε την έτοιμη συνάρτηση `sample` της R.

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-Ex2.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-Ex2.pdf.

- Παρακαλώ χρησιμοποιήστε **ένα εξώφυλλο στο pdf αρχείο σας**, στο οποίο να αναγράφεται ο τίτλος της εργασίας (Στατιστική Συμπερασματολογία), **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 Τετάρτη 17/5/2023 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