

Στατιστικά Προγράμματα, 20-01-05, Homework #2

A. Ο επόμενος πίνακας συχνοτήτων αφορά το ναυάγιο του Τιτανικού που συνέβη στις αρχές του προηγούμενου αιώνα. Δίνεται το πλήθος των διασωθέντων ανά κατηγορία επιβατών (πλήρωμα, επιβάτες 1^{ης}, 2^{ης}, 3^{ης} θέσης) και ανά φύλο.

		SURVIVED				SURVIVED	
CLASS		no	yes	CLASS		no	yes
Females:	crew	3 + x	20	Males:	crew	670 + x	192
	first	4 + x	141		first	118 + x	62
	second	13 + x	93		second	154 + x	25
	third	106 + x	90		third	422 + x	88

Όπου x είναι το τελευταίο ψηφίο του αριθμού μητρώου σας (0-9).

1. Να κάνετε ένα ραβδόγραμμα (Barchart/stacked) που να απεικονίζει πόσοι διασώθηκαν και πόσοι όχι ανά κατηγορία επιβατών (σε κάθε ράβδο να φαίνεται και το αντίστοιχο πλήθος επιβατών). Αν η διάσωση ή όχι ενός ατόμου ήταν ανεξάρτητη από την θέση στην οποία επέβαινε ποιες θα ήταν οι αναμενόμενες συχνότητες; Να ελέγξετε αν η διάσωση ήταν ανεξάρτητη από την θέση.
2. Να κάνετε ανάλογο (του ερωτ. 1) ραβδόγραμμα διασωθέντων και μη, αυτή τη φορά ανά φύλο. Αν η διάσωση ή όχι ενός ατόμου ήταν ανεξάρτητη από το φύλο ποιες θα ήταν οι αναμενόμενες συχνότητες; Να ελέγξετε αν η διάσωση ήταν ανεξάρτητη από το φύλο. Να επαναλάβετε τα προηγούμενα (ερωτ. 2) αυτή τη φορά ξεχωριστά για τις 4 κατηγορίες επιβατών (split file)
3. Να ελέγξετε αν οι άνδρες της 3^{ης} θέσης είχαν διαφορετική «τύχη» από τους άνδρες του πληρώματος.

B. Έγιναν μετρήσεις σε ένα δείγμα 400 ενηλίκων ανδρών και γυναικών καταγράφοντας τις εξής μεταβλητές: WEIGHT: βάρος (σε kgr), HEIGHT: ύψος (σε cm), WAIST: περιφέρεια μέσης (σε cm), THIGH: περιφέρεια μηρού (σε cm), GENDER: φύλο (1 - male, 0 - female). Τα δεδομένα είναι τα ίδια με αυτά που δόθηκαν στην εργασία 1 (Homework#1) **ανάλογα με τον Α.Μ. κάθε φοιτητή**. Επιθυμούμε να βρούμε έναν τρόπο εκτίμησης του σωματικού βάρους ενός ενήλικα με βάση τις υπόλοιπες μεταβλητές.

1. Για την πρόβλεψη του βάρους εφαρμόστε το μοντέλο (M1) $WEIGHT = b_0 + b_1 \cdot HEIGHT + b_2 \cdot WAIST + b_3 \cdot THIGH + \varepsilon$:
(α) Να εκτιμήσετε τα b_i σημειακά και με δ.ε. συντελεστού 95%. Επηρεάζουν όλες οι μεταβλητές το βάρος; (β) Ποια είναι η εκτίμηση της τυπικής απόκλισης των σφαλμάτων και τι ποσοστό της μεταβλητότητας του βάρους ερμηνεύεται από το μοντέλο;
2. Στο παραπάνω μοντέλο δεν ελήφθη υπόψη το φύλο. Ας θεωρήσουμε τώρα διαφορετικό μοντέλο για τα δύο φύλα ως εξής

- για τις γυναίκες: $WEIGHT = b_0 + b_1 \cdot HEIGHT + b_2 \cdot WAIST + b_3 \cdot THIGH + \varepsilon$
- για τους άντρες: $WEIGHT = b_0 + (b_1 + \gamma_1) \cdot HEIGHT + (b_2 + \gamma_2) \cdot WAIST + (b_3 + \gamma_3) \cdot THIGH + \varepsilon$

(π.χ. γ_1 είναι η διαφορά που έχει ο συντελεστής του ύψους μεταξύ ανδρών - γυναικών κ.ο.κ). Τα δύο παραπάνω μοντέλα γράφονται ενιαία σε ένα μοντέλο (M2):

$$WEIGHT = b_0 + b_1 \cdot HEIGHT + b_2 \cdot WAIST + b_3 \cdot THIGH + \gamma_1 \cdot HEIGHT \cdot GENDER + \gamma_2 \cdot WAIST \cdot GENDER + \gamma_3 \cdot THIGH \cdot GENDER + \varepsilon$$

όπου η μεταβλητή GENDER παίρνει τις τιμές 0 (γυναίκες) και 1 (άνδρες). Εφαρμόστε το παραπάνω μοντέλο αφού πρώτα δημιουργήσετε τρεις νέες μεταβλητές $HG = HEIGHT \cdot GENDER$, $WG = WAIST \cdot GENDER$, $TG = THIGH \cdot GENDER$. Ποιες από τις HG, WG, TG είναι σημαντικές (οι συντελεστές τους δεν είναι 0); Ποιές από τις μεταβλητές HEIGHT, WAIST, THIGH επηρεάζουν το βάρος διαφορετικά στους άνδρες στις γυναίκες;

3. Εφαρμόστε το μοντέλο

$$(M3) \quad WEIGHT = b_0 + b_1 \cdot HEIGHT + b_2 \cdot WAIST + b_3 \cdot THIGH + \gamma_2 \cdot WAIST \cdot GENDER + \gamma_3 \cdot THIGH \cdot GENDER + \varepsilon$$

και δώστε τις εκτιμήσεις των παραμέτρων του. Ποιό από τα μοντέλα (M1), (M2), (M3) είναι καλύτερο και γιατί;

4. Με βάση το μοντέλο (M3): (α) Δώστε προβλέψεις (σημειακές και μέσω διαστήματος) για το βάρος ενός άνδρα και μιας γυναίκας με ύψος 170cm, περιφ. μέσης 85 cm και περιφ. μηρού 60cm. (β) Η αύξηση της περιφέρειας της μέσης επηρεάζει περισσότερο το βάρος ενός άνδρα ή το βάρος μιας γυναίκας; Απαντήστε το ίδιο για την περιφέρεια του μηρού. (γ) Να γίνει έλεγχος ορθότητας του μοντέλου: Εξετάστε αν τα τυποποιημένα κατάλοιπα προέρχονται από κανονική κατανομή (ιστόγραμμα, Q-Q plot και K-S τεστ). Υπάρχουν έκτροπες παρατηρήσεις; Εξετάστε (μέσω κατάλληλου γραφήματος) αν υπάρχει σχέση μεταξύ των προσαρμοσμένων τιμών των Y_i και των τυποποιημένων καταλοίπων.

Γ. Εκλέχθηκαν τυχαία 40 δείγματα από πέντε τύπους αυτοκινήτων 1,2,3,4,5 (8 δείγματα για κάθε τύπο) και μετρήθηκε η κατανάλωση βενζίνης Y σε κάθε περίπτωση (κάτω από τις ίδιες συνθήκες κυκλοφορίας). Τα αποτελέσματα (σε Km/γαλόνι) ήταν:

Τύπος 1	13,0+x/2	15,2	21,2	21,4	21,7	17,7	19,3	16,5
Τύπος 2	25,4+x/2	27,4	29,1	30,1	25,6	23,5	29,2	29,8
Τύπος 3	16,6+x/2	23,0	20,6	19,9	22,0	25,2	16,7	25,2
Τύπος 4	27,6+x/2	24,4	28,9	32,8	20,3	27,8	22,4	23,6
Τύπος 5	31,8+x/2	32,3	31,1	27,9	30,2	35,7	32,1	38,5

Όπου x είναι το τελευταίο ψηφίο του αριθμού μητρώου σας (0-9).

Επιθυμούμε να διερευνήσουμε αν υπάρχουν διαφορές στην μέση κατανάλωση μεταξύ αυτών των 5 τύπων αυτοκινήτων. Εφαρμόστε το μοντέλο $Y_{ij} = \mu_i + \varepsilon_{ij}$, $i = 1, 2, 3, 4, 5$, $j = 1, 2, \dots, 8$ όπου μ_i είναι η μέση κατανάλωση του i -τύπου αυτοκινήτου.

1. Κατασκευάστε ένα κοινό scatterplot και ένα κοινό Boxplot της κατανάλωσης βενζίνης Y ανά τύπο αυτοκινήτου.
2. Εκτιμήστε τα μ_i σημειακά και με δ.ε. 95% και εξετάστε την υπόθεση $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ με H_1 : διαφορετικά. Να κατασκευάσετε και ένα γράφημα με τα δ.ε. των μ_i .
3. Να δοθούν ταυτόχρονα δ.ε. για τις διαφορές των μέσων ανά δύο με την μέθοδο Tukey. Να γίνει ομαδοποίηση των μέσων $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$ με πιθανότητα σφάλματος 5%.
4. Να ελέγξετε αν οι διασπορές των παρατηρήσεων είναι ίδιες στις περιπτώσεις 1, 2, 3, 4, 5 με $\alpha = 0.05$.

Όλοι οι έλεγχοι να γίνουν σε ε.σ. $\alpha = 5\%$