

Στατιστικά Προγράμματα, 21-12-05, Homework #2

A. Ο επόμενος πίνακας συχνοτήτων αφορά την εξέταση του μαθήματος «Στατιστικά Προγράμματα» (εξεταστική περίοδος Φεβρουαρίου 2005). Δίνεται το πλήθος των φοιτητών και των φοιτητριών ανά εργαστηριακό διδάσκοντα που παρακολούθησαν (Διδάσκων 1,2,3,4), ανά πλήθος απουσιών στα εργαστήρια (0-1, >1 απουσίες) και ανά τελικό βαθμό (0-4, 5-8, 9-10, δεν εξετάστηκε). Σε παρενθέσεις δίνονται πόσοι από τους φοιτητές – φοιτήτριες της κάθε κατηγορίας παρακολούθησαν σχεδόν ανελλιπώς και τα μαθήματα μεθοδολογίας (π.χ. «6 (1)» σημαίνει ότι από τα 6 άτομα της κατηγορίας, 1 παρακολούθησε και τα μαθήματα μεθοδολογίας).

	Φοιτήτριες								Φοιτητές							
	Διδάσκων 1		Διδάσκων 2		Διδάσκων 3		Διδάσκων 4		Διδάσκων 1		Διδάσκων 2		Διδάσκων 3		Διδάσκων 4	
	Απουσίες		Απουσίες		Απουσίες		Απουσίες		Απουσίες		Απουσίες		Απουσίες		Απουσίες	
	0-1	>1	0-1	>1	0-1	>1	0-1	>1	0-1	>1	0-1	>1	0-1	>1	0-1	>1
0 - 4	1	5+x ₂	7	6 (1)	5+x ₄	3	1	5	4	3	2+x ₇	1	5 (2)	4	3+x ₉	4
Βαθμός 5 - 8	4+x ₁	2	3 (1)	2+x ₃	4 (1)	1	10+x ₅	2	7+x ₆	4	7 (1)	1	6+x ₈	1	3	4+x ₀
9-10	1 (1)	1	5 (3)	0	6 (2)	1	2 (1)	1	2	1	2 (1)	0	2 (1)	0	1	0
Δεν εξετ.	0	2	0	2	0	2	1	4	1	0	0	1	1	3	0	2

(οι φοιτητές – τρεις με τελευταίο ψηφίο Α.Μ. το i να θέσουν όπου $x_i = 1$ και όλα τα άλλα $x_j = 0$).

Αφού εισάγετε τα δεδομένα στο SPSS: (Χρησιμοποιήστε όπου χρειάζεται value labels) να δώσετε σε έναν πίνακα του SPSS τον πίνακα συχνοτήτων που να περιλαμβάνει όλες τις μεταβλητές (περίπου όμοιο με τον παραπάνω). Στη συνέχεια θεωρήστε την τιμή 4: «δεν εξετ.» στο βαθμό, ως missing value.

1. (α) Να κάνετε ένα πίνακα και ένα ραβδόγραμμα (Barchart/stacked) που να απεικονίζει πόσα άτομα πήραν βαθμούς 0-4, 5-8, 9-10 ανά διδάσκοντα (στον πίνακα να φαίνονται και το αντίστοιχα ποσοστά ανά διδάσκοντα). Αν ο βαθμός ενός ατόμου ήταν ανεξάρτητος από τον εργαστ. διδάσκοντα που παρακολουθούσε ποιες θα ήταν οι αντίστοιχες αναμενόμενες συχνότητες; Ο βαθμός ήταν ανεξάρτητος από τον διδάσκοντα;

(β). Να κάνετε ανάλογο (του ερωτ. (α)) πίνακα και ραβδόγραμμα των βαθμών, αυτή τη φορά ανά φύλο (φοιτητές – φοιτήτριες). Να ελέγξετε αν ο βαθμός είναι ανεξάρτητος από το φύλο.

(γ). Να κάνετε ανάλογο (του ερωτ. (α)) πίνακα και ραβδόγραμμα των βαθμών, αυτή τη φορά ανά κατηγορία απουσιών. Να ελέγξετε αν ο βαθμός είναι ανεξάρτητος από το πλήθος απουσιών.

(δ). Να κάνετε ανάλογο (του ερωτ. (α))) πίνακα και ραβδόγραμμα των βαθμών, αυτή τη φορά ανά τα άτομα που παρακολουθούσαν ή όχι τα μαθήματα μεθοδολογίας. Να ελέγξετε αν ο βαθμός είναι ανεξάρτητος από την παρακολούθηση των μαθημάτων αυτών.

2. (α) Τα άτομα που παρακολουθούσαν τα μαθήματα μεθοδολογίας προέρχονταν ομοιόμορφα και από τα 4 τμήματα (4 διδάσκοντες);

(β) Αν θεωρηθεί ότι τα άτομα επέλεξαν μόνα τους το εργαστηριακό τμήμα που θα παρακολουθούσαν, οι φοιτήτριες είχαν κάποια ιδιαίτερη προτίμηση σε κάποιο από τα 4 τμήματα (4 διδάσκοντες);

(γ) Να ελέγξετε αν οι φοιτητές που παρακολουθούσαν τα μαθήματα μεθοδολογίας τα πήγαν καλύτερα από τους φοιτητές που δεν τα παρακολουθούσαν.

B. Έγιναν μετρήσεις σε 60 διαφορετικά είδη θηλαστικών και για κάθε είδος καταγράφηκαν οι ακόλουθες μεταβλητές:

- | | |
|--|--|
| 1. SPECIES: είδος, | 2. BODYWT: σωματικό βάρος (kg), |
| 3. BRAIWT: βάρος εγκεφάλου (gr), | 4. TOTALSL: συνολική διάρκεια ύπνου (ώρες ανα 24ωρο), |
| 5. LIFESPAN: χρόνος ζωής (έτη), | 6. GESTATION: Χρόνος κύησης (σε ημέρες) |
| 7. PREDATOR: δείκτης αρπακτικού
(1: αρπακτικό έως 5: μη αρπακτικό), | 8. DANGER: δείκτης κινδύνου που διατρέχει από άλλα αρπακτικά
(1: μικρότερος έως 5: μεγαλύτερος κίνδυνος). |

(Επειδή υπάρχουν αρκετά missing values χρησιμοποιήστε σε όσες αναλύσεις χρειάζεται την επιλογή exclude cases pairwise)

- Τα δεδομένα υπάρχουν στην Ιστοσελίδα του μαθήματος σε αρχεία του SPSS. Σε κάθε φοιτητή – φοιτήτρια αντιστοιχεί ένα σετ δεδομένων ανάλογο με το τελευταίο ψηφίο του Α.Μ. του.

(Σε κάθε φοιτητή-τρια αντιστοιχούν 60 τυχαία επιλεγμένα cases από ένα σύνολο 62 cases που προέρχονται από το άρθρο των Allison, T., and Cicchetti, D. V. (1976). Sleep in mammals: ecological and constitutional correlates. *Science* **194**, 732-734)

1. Επιθυμούμε να διερευνήσουμε τη σχέση μεταξύ σωματικού και εγκεφαλικού βάρους. Εφαρμόστε το μοντέλο

$$(M1) \quad \text{BRAIWT} = b_0 + b_1 \cdot \text{BODYWT} + \varepsilon$$

(α) Να κάνετε το διάγραμμα διασποράς των παρατηρήσεων στο επίπεδο μαζί με την εκτιμημένη ευθεία γραμμικής παλινδρόμησης και να εκτιμήσετε τα b_i σημειακά. Τι ποσοστό της μεταβλητότητας του σωματικού βάρους ερμηνεύεται από το μοντέλο; Από τον έλεγχο των καταλοίπων φαίνεται ορθό το μοντέλο αυτό;

(β) Μήπως το μοντέλο (Log είναι ο φυσικός λογάριθμος)

$$(M2) \quad \text{Log}(\text{BRAIWT}) = b_0 + b_1 \cdot \text{Log}(\text{BODYWT}) + \varepsilon$$

είναι καλύτερο και γιατί; Να κάνετε και πάλι το διάγραμμα διασποράς των παρατηρήσεων μαζί με την εκτιμημένη ευθεία γραμμικής παλινδρόμησης. Να εκτιμήσετε τα b_i σημειακά και με δ.ε. συντελεστού 95%. Τι ποσοστό της μεταβλητότητας του βάρους ερμηνεύεται τώρα από το μοντέλο; Υπάρχουν ασυνήθιστες ή έκτροπες παρατηρήσεις και σε ποια είδη θηλαστικών αντιστοιχούν;

(γ) Στο (M2): Αν ένα είδος έχει 50% περισσότερο σωματικό βάρος από ένα άλλο, πόσο % (περίπου) μεγαλύτερο βάρος εγκεφάλου θα έχει; Αν ένα τυχαίο θηλαστικό έχει σωματικό βάρος 5 κιλά, μεταξύ ποιών ορίων θα είναι το βάρος του εγκεφάλου του; (95%).

(δ) Ονομάστε C την μεταβλητή των studentized καταλοίπων του μοντέλου (M2). Μπορεί η C να θεωρηθεί ότι εκφράζει ένα (χοντροειδή) δείκτη «ευφυΐας» κάθε είδους και γιατί; Κατατάξτε σε έναν πίνακα τα είδη θηλαστικών σε σχέση με αυτό τον δείκτη.

2. Επιθυμούμε να διερευνήσουμε πως επηρεάζεται η συνολική διάρκεια ύπνου ενός θηλαστικού από τις υπόλοιπες μεταβλητές. (α) Να κάνετε έναν πίνακα με όλα τα διαγράμματα διασποράς των μεταβλητών

{Log(TOTALSL), Log(BODYWT), Log(BRAIWT), Log(LIFESPAN), Log(GESTATION), PREDATOR, DANGER}

ανά δύο (Scatterplot/matrix). Με ποιες μεταβλητές φαίνεται να σχετίζεται η μεταβλητή Log(TOTALSL) από τα διαγράμματα; Στη συνέχεια εφαρμόστε το μοντέλο

$$(M3) \quad \text{Log}(\text{TOTALSL}) = b_0 + b_1 \cdot \text{Log}(\text{BODYWT}) + b_2 \cdot \text{Log}(\text{BRAIWT}) + b_3 \cdot \text{Log}(\text{LIFESPAN}) + b_4 \cdot \text{Log}(\text{GESTATION}) + b_5 \cdot \text{PREDATOR} + b_6 \cdot \text{DANGER} + \varepsilon$$

Να εκτιμήσετε τα b_i σημειακά και με δ.ε. συντελεστού 95%. Ποιες από τις ερμηνευτικές μεταβλητές επηρεάζουν τον συνολικό ύπνο; Ποια είναι η εκτίμηση της τυπικής απόκλισης των σφαλμάτων και τι ποσοστό της μεταβλητότητας της διάρκειας του ύπνου ερμηνεύεται από το μοντέλο; Ποιο μοντέλο θα θεωρούσατε καλύτερο, το παραπάνω (M3) ή το ακόλουθο

$$(M4) \quad \text{Log}(\text{TOTALSL}) = b_0 + b_1 \cdot \text{Log}(\text{BODYWT}) + b_2 \cdot \text{Log}(\text{GESTATION}) + b_3 \cdot \text{DANGER} + \varepsilon$$

και γιατί; Ερμηνεύστε το μοντέλο (M4) και με βάση το μοντέλο αυτό:

(β) Δώστε προβλέψεις (σημειακές και μέσω διαστήματος) για τη διάρκεια του ύπνου ενός θηλαστικού με διάρκεια κύησης 3 μήνες, σωματικό βάρος 2kg και δείκτη κινδύνου 5.

(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.

(δ) Ποια θηλαστικά παρουσιάζουν ασυνήθιστα μεγάλη διάρκεια ύπνου;

Γ. Η χρονική διακριτική συχνότητα (critical flicker frequency, CFF) είναι η μέγιστη συχνότητα (εκφρασμένη σε Hz ή cycles/sec) στην οποία μπορεί το οπτικό σύστημα ενός ατόμου να διακρίνει ότι «αναβοσβήνει» ένα οπτικό σήμα. Αν ένα οπτικό σήμα «αναβοσβήνει» με μεγαλύτερη συχνότητα από το CFF, το άτομο το αντιλαμβάνεται σαν ένα συνεχές σήμα (χωρίς «flickering»). Στα πλαίσια μιας έρευνας καταγράφηκε ο δείκτης CFF σε 19 άτομα. Επίσης καταγράφηκε και το χρώμα του ματιού του κάθε ατόμου. (Τα δεδομένα προέρχονται από το άρθρο: The effect of iris color on critical flicker frequency. *Journal of General Psychology* (1973), 91-95)

CFF (Brown)	26.8	27.9	23.7	25	26.3	24.8	25.7	24.5+ x/5
CFF (Green)	26.4	24.2	28	26.9	29.1+ x/5			
CFF (Blue)	25.7	27.2	29.9	28.5	29.4	28.3+ x/5		

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

Επιθυμούμε να διερευνήσουμε αν υπάρχουν διαφορές στη μέση χρονική διακριτική συχνότητα μεταξύ ατόμων με διαφορετικό χρώμα της ίριδας του ματιού. Εφαρμόστε το μοντέλο $Y_{ij} = \mu_i + \varepsilon_{ij}$, $i = 1, 2, 3$, όπου μ_i είναι η μέση χρονική διακριτική συχνότητα ατόμων με i-χρώμα ίριδας (1:Brown, 2:Green, 3:Blue)

1. Κατασκευάστε ένα κοινό scatterplot και ένα κοινό Boxplot της CFF ανά χρώμα ίριδας.

2. Εκτιμήστε τα μ_i σημειακά και με δ.ε. 95% και εξετάστε την υπόθεση $H_0: \mu_1 = \mu_2 = \mu_3$ με H_1 : διαφορετικά. Να κατασκευάσετε και ένα γράφημα με τα δ.ε. των μ_i (error bar).

3. Να δοθούν ταυτόχρονα δ.ε. για τις διαφορές των μέσων ανά δύο με την μέθοδο Tukey. Να γίνει ομαδοποίηση των μέσων μ_1, μ_2, μ_3 με πιθανότητα σφάλματος 5%.

4. Να ελέγξετε αν οι διασπορές των παρατηρήσεων είναι ίδιες στις ομάδες 1, 2, 3.

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