

Ανάλυση Δεδομένων με R/Y

ΘΕΜΑ 1 (3 μονάδες):

A) Η συνάρτηση μάζας πιθανότητας της κατανομής Poisson είναι η

$$P(X=x) = \frac{e^{-\lambda} \lambda^x}{x!}, x=0,1,\dots, \lambda > 0.$$

Επειδή όμως χρειάζεται να υπολογιστεί το παραγοντικό στον παρονομαστή, ο παραπάνω ορισμός δεν είναι ιδιαίτερα εύχρηστος. Για αυτό το λόγο στην πράξη χρησιμοποιούμε αναδρομικό τύπο και συγκεκριμένα, υπολογίζουμε την $P(X=0) = e^{-\lambda}$ και στη συνέχεια οι υπόλοιπες προκύπτουν ως

$$P(X=x) = P(X=x-1) \frac{\lambda}{x}, x=1,2,\dots$$

Να γράψετε μια συνάρτηση στην R που να έχει ως παραμέτρους εισόδου την τιμή της παραμέτρου λ , και την τιμή του x για το οποίο θέλετε να υπολογίσετε την πιθανότητα και η οποία να επιστρέφει ως αποτέλεσμα την συνάρτηση μάζας πιθανότητας στο σημείο x. Αν η τιμή εισόδου x δεν είναι φυσικός αριθμός καθώς και αν η τιμή εισόδου λ δεν είναι μεγαλύτερη του μηδενός η συνάρτησή σας θα πρέπει να επιστρέφει ένα μήνυμα λάθους.

B) Έστω η τ.μ. $X \sim \text{Γάμμα}(4,5)$. Με ποιες εντολές της R θα υπολογίζατε:

- Την πιθανότητα $P(X > 1)$.
- Την τιμή a τέτοια ώστε $P(X \leq a) = 0.5$.
- Την τιμή της σ.π.π. στο σημείο 3.

ΘΕΜΑ 2 (1.9 μονάδες):

Μια εταιρεία εμπορίας υγρών καυσίμων εισήγαγε στην αγορά το τελευταίο εξάμηνο ένα νέο προϊόν βενζίνης υψηλών προδιαγραφών. Το τμήμα πωλήσεων της εταιρείας προκειμένου να καθορίσει τη συμπεριφορά των πωλήσεων του νέου προϊόντος σε σχέση με την τιμή του και τις δαπάνες διαφήμισης, συνέλεξε δεδομένα για τις τελευταίες 16 εβδομάδες για τις παρακάτω μεταβλητές.

X_1 : εβδομαδιαίες πωλήσεις του νέου προϊόντος σε χιλιάδες λίτρα.

X_2 : μέση εβδομαδιαία τιμή του προϊόντος σε ευρώ.

X_3 : εβδομαδιαίες δαπάνες διαφήμισης σε χιλιάδες ευρώ.

- Έστω ότι τα δεδομένα, βρίσκονται υπό μορφή ενός πίνακα 16×3 στο αρχείο "data.txt". Με ποιον τρόπο μπορείτε να εισάγετε τα δεδομένα στην R και να δημιουργήσετε ένα πλαίσιο δεδομένων με ονόματα στις 3 μεταβλητές;
- Με ποιους τρόπους (αριθμητικούς και γραφικούς) και ποιες εντολές της R θα περιγράφατε τις τιμές των μεταβλητών σας;
- Τι είδους στατιστική ανάλυση θα εφαρμόζατε και με ποιες εντολές στην R;
- Ποιες προϋποθέσεις θα ελέγχατε για την στατιστική ανάλυση του παραπάνω υποερωτήματος και με ποιες εντολές στην R;

ΘΕΜΑ 3 (3.6 μονάδες):

Μια τράπεζα θέλει να φτιάξει ένα μοντέλο για να προβλέψει το συνολικό ποσό που κάνουν ανάληψη οι πελάτες της το Σαββατοκύριακο από τα μηχανήματα αυτόματων συναλλαγών (ATM) έτσι ώστε να εφοδιάζει με το σωστό ποσό το κάθε μηχάνημα. Τα διαθέσιμα δεδομένα έχουν την ακόλουθη μορφή:

ATM	Amount	Homeval	Location	Hval_cat
1	120	225	1	4
2	99	170	0	2
3
15	112	210	0	3

ATM: ο κωδικός του μηχανήματος αυτόματης συναλλαγής.

Amount: το συνολικό ποσό ανάληψης το Σαββατοκύριακο (σε χιλιάδες €).

Homeval: η διάμεσος της αξίας των σπιτιών στη περιοχή (σε χιλιάδες €).

Location: αν το ATM βρίσκεται σε Εμπορικό κέντρο ή όχι (0 = όχι, 1 = ναι).

Hval_cat: Κατηγοριοποίηση της Homeval με κατηγορίες (1 = χαμηλή αξία, 2 = μέτρια αξία, 3 = υψηλή αξία, 4 = πολύ υψηλή αξία).

Χρησιμοποιώντας τα παραπάνω δεδομένα εξετάσαμε αν το ποσό ανάληψης το Σαββατοκύριακο έχει σχέση με τις κατηγορίες της αξίας των σπιτιών και με τα μηχανήματα αυτόματων συναλλαγών βρίσκονται ή όχι σε εμπορικό κέντρο με βάση το παρακάτω γενικό γραμμικό μοντέλο:

$$E(\text{Amount} | \text{Location1}, \text{Hval_cat2}, \text{Hval_cat3}, \text{Hval_cat4}) = \alpha + \beta_1 \text{Location1} + \beta_2 \text{Hval_cat2} + \beta_3 \text{Hval_cat3} + \beta_4 \text{Hval_cat4}.$$

όπου

$$\text{Location1} = \begin{cases} 1, & \text{ATM σε εμπορικό κέντρο} \\ 0, & \text{διαφορετικά} \end{cases}, \quad \text{Hval_cat2} = \begin{cases} 1, & \text{μέτρια διάμεση αξία σπιτιών} \\ 0, & \text{διαφορετικά} \end{cases}$$

$$\text{Hval_cat3} = \begin{cases} 1, & \text{υψηλή διάμεση αξία σπιτιών} \\ 0, & \text{διαφορετικά} \end{cases}, \quad \text{Hval_cat4} = \begin{cases} 1, & \text{πολύ υψηλή διάμεση αξία σπιτιών} \\ 0, & \text{διαφορετικά} \end{cases}$$

- Με ποια εντολή της R θα προσαρμόσατε το παραπάνω γενικό γραμμικό μοντέλο;
- Ποιες είναι οι κατηγορίες αναφοράς στις εικονικές σας μεταβλητές;
- Εξηγήστε πλήρως τα παρακάτω αποτελέσματα που παίρνετε από την R ύστερα από την προσαρμογή του παραπάνω γενικού γραμμικού μοντέλου.

Residuals:					
Min	1Q	Median	3Q	Max	
-5.5682	-2.7992	0.3333	2.7008	5.4318	
 Coefficients:					
(Intercept)	83.932	2.462	34.091	1.12e-11 ***	
LocationI	1.636	2.543	0.644	0.534352	
Hva_cat2	14.159	3.049	4.644	0.000916 ***	
Hva_cat3	28.500	2.982	9.559	2.40e-06 ***	
Hva_cat4	38.098	3.462	11.003	6.57e-07 ***	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '+' 1					
 Residual standard error: 4.217 on 10 degrees of freedom					
Multiple R-squared: 0.946, Adjusted R-squared: 0.9243					
F-statistic: 43.76 on 4 and 10 DF, p-value: 2.64e-06					

- iv) Με βάση τα παραπάνω αποτελέσματα ερμηνεύστε τους εκτιμητές των συντελεστών του γενικού γραμμικού μοντέλου.
- v) Να εκτιμήσετε το συνολικό ποσό ανάληψης το Σαββατοκύριακο (σε χιλιάδες €) μηχανήματος αυτόματης συναλλαγής που δεν βρίσκεται σε εμπορικό κέντρο και η διάμεσος της αξίας των σπιτιών στη περιοχή είναι υψηλή.
- vi) Να συνοψίσετε τα συμπεράσματα της παραπάνω ανάλυσης.

ΘΕΜΑ 4 (1.5 μονάδες):

Τα παρακάτω δεδομένα αφορούν τα φορτία θραύσης (σε tN/cm²) συνθετικών νημάτων δύο τύπων.

Tύπου A	1.2	0.3	0.8	0.5	0.4	1.3	1.4	
Tύπου B	1.6	1.5	1.1	1.0	1.8	1.7	0.9	0.7

Θέλουμε να ελέγξουμε σε ε.σ. 5%, την υπόθεση ότι οι δύο τύποι νημάτων έχουν την ίδια μέση αντοχή με εναλλακτική ότι οι μέσες τιμές είναι διαφορετικές.

- i) Τι είδους στατιστική ανάλυση θα εφαρμόζετε και με ποιες εντολές στην R;
- ii) Ποιες προϋποθέσεις θα ελέγχατε για την στατιστική ανάλυση του παραπάνω υποερωτήματος και με ποιες εντολές στην R;
- iii) Αν δεν ισχυαν οι παραπάνω προϋποθέσεις ποιον έλεγχο θα εφαρμόζετε και με ποια εντολή στην R;

Διάρκεια Εξέτασης: 2 ½ h

EΥΧΟΜΑΣΤΕ ΕΠΙΤΥΧΙΑ!