ΜΑΘΗΜΑ: "Πληροφορική με Εφαρμογές Στατιστικής"
ΕΡΓΑΣΤΗΡΙΟ 13: Επανάληψη

Στη διάθεσή σας έχετε τα δεδομένα του αρχείου lab13.ods (δείτε Berk and Carey (2000), dataset CALC.xls, σελ. 324). Τα δεδομένα αφορούν τις επιδόσεις πρωτοετών φοιτητών/φοιτητριών στον Απειροστικό Λογισμό (calculus). Συγκεκριμένα:

X1: Τιμή 0 αν δεν είχε παρακολουθήσει Απειροστικό Λογισμό στο λύκειο, 1 αν είχε παρακολουθήσει.

X2: Επίδοση στο τεστ ACT

X3: Επίδοση σε εισαγωγικό τεστ κατάταξης στην Άλγεβρα

X4: Βαθμός επίδοσης στην Άλγεβρα λυκείου

X5: Κατάταξη στο Λύκειο

X6: Φύλο

X7: Φύλο (κωδικοποιημένο) ως 0 για Γυναίκες, 1 για Άνδρες

X8: Επίδοση στον Απειροστικό Λογισμό

1. Χρησιμοποιήστε τις τιμές της X8 και υπολογίστε τα παρακάτω μέτρα: Δειγματική μέση τιμή, δειγματική τυπική απόκλιση, δειγματικός συντελεστής ασυμμετρίας $\hat{\beta}_1$, δειγματικός συντελεστής κύρτωσης $\hat{\beta}_2$. Χρησιμοποιήστε τις συναρτήσεις του CALC. Ερμηνεύστε τα αποτελέσματα των δεικτών $\hat{\beta}_1$ και $\hat{\beta}_2$ σχετικά με την κατανομή των δεδομένων.

2. Χρησιμοποιήστε τις τιμές της X8 και υπολογίστε τις τιμές των παρακάτω (δειγματικών) ποσοστιαίων σημείων: P_5, P_{25}, P_{50}, P_{75}, P_{95}. Χρησιμοποιήστε τις συναρτήσεις του CALC. Υπολογίστε την τιμή του δειγματικού ενδοτεταρτημοριακού εύρους IQR.

3. Χρησιμοποιήστε τις τιμές της X8 και υπολογίστε την τιμή του δείκτη Gini, τη δειγματική μέση απόλυτη απόκλιση MAD και το συντελεστή μεταβλητότητας. Είναι ομοιογενές το δείγμα; Όπου υπάρχει έτοιμη συνάρτηση στο CALC για τα παραπάνω μέτρα, να τη χρησιμοποιήσετε.

4. Αν έπρεπε να ομαδοποιήσετε σε κλάσεις τα δεδομένα της X8, πόσες θα χρησιμοποιούσατε; Να αιτιολογήσετε την απάντησή σας. Αν έπρεπε να πάρετε σταθερό πλάτος για κάθε κλάση, τι θα προτείνατε; Αφού απαντήσετε στα προηγούμενα δύο ερωτήματα, να κάνετε την ομαδοποίηση των τιμών της X8 χρησιμοποιώντας ως άνω όριο κάθε κλάσης (bins) τις τιμές 50, 60, 70, 80, 90, 100. Με βάση αυτή την ομαδοποίηση, ποια είναι η επικρατούσα κλάση; Για την ομαδοποίηση, χρησιμοποιήστε τη συνάρτηση FREQUENCY.

5. Ποιο είναι το κατάλληλο διάγραμμα για τη γραφική αναπαράσταση της κατανομής των τιμών των μεταβλητών X2, X3, X4 και X8; Ποιο είναι αντίστοιχα το κατάλληλο διάγραμμα για τις μεταβλητές X1 και X6 (ή X7); Να δώσετε τον πίνακα διπλής εισόδου 2×2 για τις μεταβλητές X1 και X6 (ή X7).

6. Να υπολογιστεί και να ερμηνευτεί ο δειγματικός συντελεστής συσχέτισης r μεταξύ των μεταβλητών X2 και X8. Στη συνέχεια, να βρείτε τις τιμές του σταθερού όρου και της κλίσης της προσαρμοσμένης ευθείας γραμμικής παλινδρόμηση $\hat{y} = \hat{a} + \hat{b}X$, όπου X είναι οι τιμές της X2 και Y είναι οι τιμές της X8. Ποιο είναι το κατάλληλο διάγραμμα για να αναπαραστήσετε στο επίπεδο τις τιμές των ζευγών (X, Y);

7. Χρησιμοποιήστε Συγκεντρωτικούς Πίνακες (Pivot Tables) και βρείτε: Τη μέση επίδοση στο τεστ ACT ως προς τα 2 φύλα, τη μέση επίδοση στο εισαγωγικό τεστ κατάταξης στην Άλγεβρα ως προς τα 2 φύλα, τη μέση επίδοση στον Απειροστικό Λογισμό ως προς το αν είχε παρακολουθήσει ο/η φοιτητής/τρια Απειροστικό Λογισμό στο Λύκειο.

8. Να υπολογίσετε την τάξη (rank) για κάθε φοιτητή/τρια του δείγματος, με βάση τις τιμές της X8. Στη συνέχεια, χρησιμοποιήστε τη συνάρτηση sumif() και βρείτε το άθροισμα των τάξεων μόνο για τους φοιτητές. Χρησιμοποιήστε τη συνάρτηση averageif() και βρείτε το μέσο όρο των τάξεων μόνο για τις φοιτήτριες.

9. Κωδικοποιήστε τις τιμές της X2 ως εξής: LOW αν X2 < 20, AVERAGE αν $20 \leq X2 < 30$, EXCELLENT αν $X2 \geq 30$. Ονομάστε τη νέα μεταβλητή X2cat. Στη συνέχεια να γίνει το ομαδοποιημένο ραβδόγραμμα σχετικών συχνοτήτων της X2cat ως προς το φύλο (μεταβλητή X6 ή X7) και το αντίστοιχο στοιβαγμένο ραβδόγραμμα %. Επαναλάβετε αλλά αυτή τη φορά να γίνει το ομαδοποιημένο ραβδόγραμμα σχετικών συχνοτήτων της X6 (ή της X7) ως προς τα 3 επίπεδα της X2cat. Να γίνει και το αντίστοιχο στοιβαγμένο ραβδόγραμμα %.

10. Να γίνει το διάγραμμα φυσαλίδας (Bubble chart) των X2 (στον άξονα X), X8 (στον άξονα Y), χρησιμοποιώντας για το μέγεθος της φυσαλίδας τη μεταβλητή X5.

11. Να βρεθεί το πλήθος των μοναδικών τιμών της X4 (μπορείτε να το κάνετε με το AutoFilter). Στη συνέχεια, κατασκευάστε τον πίνακα συχνοτήτων για τις τιμές της X4.

12. Να βρεθεί το ποσοστό των μελών του δείγματος με βαθμό στον Απειροστικό Λογισμό μεγαλύτερο του 70 (δηλ. X8>70) και επίδοση στο ACT τουλάχιστον 30 (δηλ. $X2 \geq 30$). Ποιο είναι το ποσοστό των Ανδρών και ποιο το ποσοστό των Γυναικών στο παραπάνω υποσύνολο του δείγματος;

13. Υπολογίστε τον παρακάτω συντελεστή συσχέτισης για τις τιμές των X2, X3: Αρχικά βρείτε τις τάξεις των τιμών στη X2 και στη X3, έστω αυτές $r(X_2)$, $r(X_3)$. Στη συνέχεια, βρείτε το μέσο όρο των τάξεων για τις τιμές των X2, X3, έστω αυτοί $\overline{r(X_2)}$ και $\overline{r(X_3)}$**,** αντίστοιχα. Στη συνέχεια, υπολογίστε την τιμή της παρακάτω ποσότητας (η οποία είναι γνωστή και ως συντελεστής συσχέτισης του Spearman)

$$r_S = \frac{\sum_{i=1}^{n}\left(r(X_{2,i})-\overline{r(X_2)}\right)\left(r(X_{3,i})-\overline{r(X_3)}\right)}{\sqrt{\sum_{i=1}^{n}\left(r(X_{2,i})-\overline{r(X_2)}\right)^2}\sqrt{\sum_{i=1}^{n}\left(r(X_{3,i})-\overline{r(X_3)}\right)^2}}.$$

14. Να βρεθεί το άθροισμα των τετραγωνικών αποκλίσεων των τιμών της X8 από τη μέση τιμή τους (π.χ. χρησιμοποιήστε τη συνάρτηση DEVSQ()). Υπολογίστε το πλήθος των Ανδρών και το πλήθος των Γυναικών στο δείγμα (π.χ. με τη συνάρτηση COUNTIF()). Βρείτε τη μέση τιμή της X8 μόνο για τους Άνδρες και μόνο για τις Γυναίκες (π.χ. με τη συνάρτηση AVERAGEIF()). Κάντε μια νέα μεταβλητή, π.χ. X9 η οποία θα έχει την τιμή $\left(X8-\bar{X}_{8,A}\right)^2$ αν το άτομο είναι φοιτητής, διαφορετικά $\left(X8-\bar{X}_{8,\Gamma}\right)^2$ αν το άτομο είναι φοιτήτρια. Στη συνέχεια, να βρείτε τα αθροίσματα των παραπάνω τετραγωνικών αποκλίσεων, χωριστά για Άνδρες και χωριστά για γυναίκες (π.χ. με τη συνάρτηση SUMIF()). Τέλος, να επιβεβαιώσετε την παρακάτω ταυτότητα για τις τιμές της X8 ως προς τα φύλα

$$\sum_{j=1}^{\kappa}\sum_{i=1}^{n_j}\left(X_{ij}-\bar{\bar{X}}\right)^2 = \sum_{i=1}^{\kappa}\sum_{i=1}^{n_j}\left(X_{ij}-\bar{X}_{.j}\right)^2 + \sum_{j=1}^{\kappa} n_j\left(\bar{X}_{.j}-\bar{\bar{X}}\right)^2.$$

όπου *κ* είναι το πλήθος των διαφορετικών κατηγορικών (τα δύο φύλα, δηλαδή $\kappa = 2$), n_1 είναι το πλήθος των Γυναικών, n_2 το πλήθος των ανδρών στο δείγμα, $\bar{\bar{X}}$ είναι ο μέσος των τιμών της X8 και η ποσότητα στο αριστερό μέλος είναι αυτό που θα βρείτε με την DEVSQ. Επίσης $\bar{X}_{.1}$, $\bar{X}_{.2}$ είναι η μέση τιμή της X8 για τις γυναίκες και για τους άνδρες, αντίστοιχα (το έχετε ήδη βρει με την AVERAGEIF(), και τα έχετε χρησιμοποιήσει ως $\bar{X}_{8,\Gamma}$ και $\bar{X}_{8,A}$ στην κατασκευή της X9).

[Αποτέλεσμα: 10332.2 = 10327.70962 + 4.4903834]