

ΜΑΘΗΜΑ: “Πληροφορική με Εφαρμογές”

ΕΡΓΑΣΤΗΡΙΟ 7: Περιγραφική Στατιστική με το CALC – Ομαδοποιημένα Δεδομένα

Άσκηση 1^η

Στο αρχείο LAB6.ods (θα το βρείτε στα Έγγραφα, στο eclass του μαθήματος) δίνεται ο αριθμός των ελαττωματικών αντικειμένων σε συσκευασίες μεγέθους 8 αντικειμένων (φύλλο εργασίας OM1). Στο δείγμα εμφανίζονται οι τιμές 0, 1, ..., 8, με διαφορετική συχνότητα εμφάνισης η καθεμία (**άσκηση για το σπίτι**: δοκιμάστε να βρείτε το πλήθος των διαφορετικών τιμών της στήλης **X** ακολουθώντας τα βήματα που περιγράφονται στην άσκηση του εργαστηρίου 5).

(α) Χρησιμοποιήστε το CALC και συμπληρώστε τον παρακάτω πίνακα. Στις 3 τελευταίες στήλες, τα αποτελέσματα να δοθούν με ακρίβεια 4 δεκαδικών ψηφίων.

i	X_i	f_i	cf_i	$rf_i\%$	$rcf_i\%$	φ_i	$f_i \cdot X_i$	$f_i \cdot (X_i - \bar{x})^2$	$f_i \cdot (X_i - \bar{x})^3$	$f_i \cdot (X_i - \bar{x})^4$
0										
1										
2										
3										
4										
5										
6										
7										
8										
Σύνολο										

Υπόδειξη: Μπορείτε να το κάνετε με χρήση της COUNTIF. Ξέρετε ότι οι πιθανές τιμές στο δείγμα είναι 0, 1, 2, ..., 8. Άρα, για να βρείτε π.χ. πόσα είναι τα μηδενικά στο δείγμα, δώστε σε ένα κελί την εντολή
`=COUNTIF(B2:B251;"=0")`

Με τον ίδιο τρόπο μπορείτε να βρείτε και για τις υπόλοιπες τιμές, πόσες φορές εμφανίζεται καθεμιά από αυτές στο δείγμα.

Επίσης, μπορείτε να χρησιμοποιήσετε τη συνάρτηση FREQUENCY (ως array function). Δώστε αρχικά τις τιμές 0, 1, ..., 8 σε 9 κελιά (το ένα κάτω από το άλλο, π.χ. στα C2 – C10). Στη συνέχεια, επιλέξτε τα κελιά D2- D10, γράψτε την εντολή

$$=FREQUENCY(B2:B251;C2-C10)$$

και πατήστε CTRL + SHIFT + ENTER.

(α) Χρησιμοποιώντας τις πληροφορίες μόνο από τον πίνακα συχνοτήτων, να υπολογίσετε το δειγματικό μέσο όρο των ελαττωματικών αντικειμένων. Ο τύπος υπολογισμού είναι

$$\bar{X} = \frac{1}{n} \sum_{i=1}^k f_i X_i = \sum_{i=1}^k r f_i X_i,$$

όπου k είναι το πλήθος των διαφορετικών τιμών στο δείγμα. Εδώ το $k = 9$. Ένας γρήγορος τρόπος για να το βρείτε είναι με χρήση της συνάρτησης =SUMPRODUCT(array1;array2). Στο 1^ο array θα είναι οι τιμές των 0, 1, ..., 8 και στο 2^ο array θα είναι οι αντίστοιχες σχετικές συχνότητες $r f_1, \dots, r f_9$.

(β) Χρησιμοποιώντας ΜΟΝΟ τις πληροφορίες που σας δίνει ο παραπάνω πίνακας, να απαντήσετε στις επόμενες ερωτήσεις:

1. Πόσες συσκευασίες έχουν το πολύ 3 ελαττωματικά αντικείμενα; Ποιο το αντίστοιχο ποσοστό;
2. Πόσες συσκευασίες έχουν περισσότερα από 6 ελαττωματικά αντικείμενα; Ποιο το αντίστοιχο ποσοστό;
3. Πόσες συσκευασίες έχουν τουλάχιστον 5 ελαττωματικά αντικείμενα; Ποιο το αντίστοιχο ποσοστό;
4. Πόσες συσκευασίες έχουν λιγότερα από 2 ελαττωματικά αντικείμενα; Ποιο το αντίστοιχο ποσοστό;
5. Πόσες συσκευασίες ακριβώς 7 ελαττωματικά αντικείμενα; Ποιο το αντίστοιχο ποσοστό;

(γ) Χρησιμοποιώντας ΜΟΝΟ τις πληροφορίες που σας δίνει ο παραπάνω πίνακας, να υπολογίσετε τα παρακάτω αριθμητικά μέτρα: Δειγματική διασπορά, δειγματική τυπική απόκλιση, συντελεστής ασυμμετρίας β_1 και συντελεστής κύρτωσης β_2 .

Υπόδειξη: Να χρησιμοποιήσετε τους παρακάτω τύπους για τον υπολογισμό της διασποράς και των συντελεστών β_1, β_2

$$S^2 = \frac{1}{n-1} \sum_{i=1}^k f_i (X_i - \bar{X})^2, \quad \beta_1 = \left(\frac{1}{n} \sum_{i=1}^k f_i (X_i - \bar{X})^3 \right) / S^3, \quad \beta_2 = \left(\frac{1}{n} \sum_{i=1}^k f_i (X_i - \bar{X})^4 \right) / S^4.$$

(δ) Μπορούμε να θεωρήσουμε ότι η κατανομή των δεδομένων είναι συμμετρική; Τι μπορείτε να πείτε για την κύρτωση της κατανομής των δεδομένων;

(ε) Να υπολογίσετε τις παρακάτω τιμές για το διαθέσιμο σετ δεδομένων: Ελάχιστη παρατήρηση, 1^ο τεταρτημόριο, διάμεσος, 3^ο τεταρτημόριο, Μέγιστη παρατήρηση, Κορυφή (χρησιμοποιήστε μόνο τις πληροφορίες που σας δίνει ο πίνακας συχνοτήτων, δηλ. να μην χρησιμοποιηθούν οι αντίστοιχες συναρτήσεις του CALC).

Άσκηση 2^η

Ανοίξτε το CALC και φορτώστε το αρχείο Lab6.ods (φύλλο εργασίας OM2). Στη στήλη **AA** δίνεται ο αύξων αριθμός των δεδομένων στο αρχείο και στη στήλη **X** δίνεται η τιμή της χοληστερίνης 40 ενηλίκων. Χρησιμοποιώντας τα δεδομένα στο φύλλο εργασίας OM2 και τις αντίστοιχες συναρτήσεις του CALC, να κάνετε ομαδοποίηση των δεδομένων σε κλάσεις ίσου πλάτους (να επιλέξετε εσείς τον αριθμό των κλάσεων, σύμφωνα με τα όσα έχουμε πει) και να δώσετε τον πίνακα συχνοτήτων. Δώστε έμφαση στο πως μπορείτε να χρησιμοποιήσετε για το σκοπό αυτό τη συνάρτηση FREQUENCY.

Υπολογισμός του πλήθους k των κλάσεων, για δείγμα n παρατηρήσεων.

Sturges formula: $1 + \log_2 n$ και στρογγυλοποίηση στον μεγαλύτερο ακέραιο. Για το πλάτος της κλάσης, $c = \frac{X_{(n)} - X_{(1)}}{k}$. Το c δε στρογγυλοποιείται κατ' ανάγκη σε ακέραιο αλλά στρογγυλοποιείται προς τα πάνω.

Freedman – Diaconis rule: Πλάτος κλάσης $c = 2 \cdot \frac{IQR}{\sqrt[3]{n}}$. Στη συνέχεια, το $k = \frac{X_{(n)} - X_{(1)}}{c}$ και η στρογγυλοποίηση γίνεται στον μεγαλύτερο ακέραιο.

Ο πίνακας πρέπει να περιέχει τα εξής στοιχεία: Όρια κλάσεων (κάτω και άνω), κεντρικές τιμές y_i για κάθε κλάση, τιμές για απόλυτες και αθροιστικές συχνότητες, σχετικές και σχετικές αθροιστικές συχνότητες (οι δύο τελευταίες, σε μορφή ποσοστού). Η μορφή του να είναι (κάπως) έτσι:

i κλάση	Κάτω Όριο	Άνω Όριο	y_i	f_i	rf_i	$rf_i\%$	$rcf_i\%$
1							
2							
3							
...							
$k-1$							
k							
Σύνολο							

Χρησιμοποιώντας MONO τα στοιχεία του πίνακα συχνοτήτων, να υπολογίσετε μέση τιμή και τυπική απόκλιση των (ομαδοποιημένων) δεδομένων. Αφού υπολογίσετε τις τιμές αυτές, στη συνέχεια να υπολογίσετε μέση τιμή και τυπική απόκλιση των μη-ομαδοποιημένων δεδομένων. Παρατηρείτε διαφορές στις τιμές των μέτρων; Που μπορεί να οφείλεται αυτό; Για τον υπολογισμό της μέσης τιμής και της διασποράς στα ομαδοποιημένα δεδομένα, να χρησιμοποιήσετε τους παρακάτω τύπους

$$\bar{X} = \frac{1}{n} \sum_{i=1}^k f_i y_i, \quad S^2 = \frac{1}{n-1} \sum_{i=1}^k f_i (y_i - \bar{X})^2 = \frac{1}{n-1} (\sum_{i=1}^k f_i y_i^2 - n \bar{X}^2),$$

όπου k είναι το πλήθος των κλάσεων και y_i είναι η κεντρική τιμή της i -οστής κλάσης.

Να επαναλάβετε τα παραπάνω ερωτήματα αν είναι γνωστές οι κλάσεις για την ομαδοποίηση των δεδομένων: [170, 185), [185, 200), [200, 215), [215, 230). Παρατηρείτε διαφορές στις τιμές της μέσης τιμής και της τυπικής απόκλισης έναντι της προηγούμενης ομαδοποίησης; Υπάρχουν διαφορές αν η ομαδοποίηση γίνει στις κλάσεις (170, 185], (185, 200], (200, 215], (215, 230];

Πριν φύγετε, να αποθηκεύσετε (σε δικό σας φάκελο, είτε στο σκληρό δίσκο του Η/Υ στο εργαστήριο, είτε σε “φλασάκι”) το βιβλίο εργασίας που δημιουργήσαμε στο 6ο εργαστήριο. Να δώσετε το όνομα LAB6.ods. Ενδέχεται να το χρησιμοποιήσουμε και σε επόμενη διάλεξη.