Μετρήσεις

Χωρισμός των μεταβλητών σε κλάσεις

Αν στα δεδομένα υπάρχουν πολλές διαφορετικές τιμές το ολικό διάστημα της μεταβλητής μπορεί να διαιρεθεί σε κλάσεις, οι οποίες δεν καλύπτουν η μία την άλλη, και η ένωση τους καλύπτει το συνολικό διάστημα. Συνήθως το εύρος των κλάσεων είναι σταθερό (το ίδιο για όλες τις κλάσεις), μπορεί όμως να είναι και άνισο ανάλογα με τη φύση της μεταβλητής και του πληθυσμού.

Ειδικότερα, έστω ότι θέλουμε να οργανώσουμε σε μια κατανομή συχνοτήτων τις παρατηρήσεις $X_1, X_2,...,X_n$ μιας συνεχούς μεταβλητής X (η X μπορεί ακόμη να είναι διακριτή με μεγάλο πλήθος τιμών οπότε αντιμετωπίζεται ως συνεχής). Για το σκοπό αυτό χωρίζουμε το εύρος των δεδομένων ίσα διαδοχικά διαστήματα ή τάξεις και καταγράφουμε τη συχνότητα των παρατηρήσεων που ανήκουν σε κάθε τάξη.

Το πλήθος κ των τάξεων επιλέγεται παίρνοντας υπόψη το πλήθος των παρατηρήσεων καθώς και τον σκοπό για τον οποίο προορίζεται η κατανομή. Όταν τα δεδομένα είναι λιγότερα από 100 επιλέγουμε συνήθως μέχρι 10 τάξεις. Από 100 ως 1000 συνήθως αρκούν 10-12 τάξεις. Από 1000 ως 10000 αρκούν 13-16 τάξεις. Είναι δυνατόν να μειώσουμε τον αριθμό των τάξεων αν υπάρχουν πολλές μηδενικές συχνότητες. Ή ακόμη να τον αυξήσουμε αν η κατανομή είναι πολύ ανώμαλη. Όταν σκοπός της ομαδοποίησης των δεδομένων είναι απλώς η περιληπτική τους παρουσίαση, το πλήθος των τάξεων δεν πρέπει να ξεπερνά το 20. Να σημειωθεί ότι όταν σκοπός της ομαδοποίησης των δεδομένων είναι η περαιτέρω στατιστική τους ανάλυση είναι χρήσιμο να δοκιμάζονται διαφορετικές ομαδοποιήσεις μεταβάλλοντας το πλήθος και φυσικά και το εύρος των τάξεων.

Για να υπολογίσουμε το εύρος της κάθε τάξης διαιρούμε το συνολικό εύρος με το πλήθος των τάξεων κ και στρογγυλοποιούμε το αποτέλεσμα έτσι ώστε να προκύπτει ένας αριθμός εύχρηστος και οικείος στον μέσο αναγνώστη. Εκείνο που πρέπει πάντοτε να διευκρινίζεται είναι πού ανήκουν τα όρια του διαστήματος που ορίζει την κάθε κλάση. Ο πιο συνηθισμένος τρόπος είναι: "το κατώτερο όριο περιλαμβάνεται στην κλάση, ενώ δεν περιλαμβάνεται το ανώτερο". Σημαντική είναι και η τιμή, η οποία επιλέγεται ως κατώτερο όριο της πρώτης τάξης. Η τιμή αυτή ονομάζεται και άγκυρα και μπορεί να επηρεάσει σημαντικά την εικόνα που θα πάρουμε από την κατανομή και την γραφική της παράσταση.

Τέλος, σημειώνεται ότι όταν αναφερόμαστε σε κλάσεις ποσοτικών μεταβλητών, τότε όλες οι μονάδες της κάθε κλάσεις θεωρούνται ότι είναι συγκεντρωμένες στο μέσο του διαστήματος που έχει και την έννοια του κέντρου μάζας. Έτσι, η κεντρική τιμή κάθε τάξής μπορεί να χρησιμοποιηθεί σε περαιτέρω υπολογισμούς. Για το λόγο αυτό επιλέγοντας τα όρια των τάξεων, δηλαδή την κατώτερη και την ανώτερη τιμή για κάθε τάξη φροντίζουμε να προκύπτουν εύχρηστες κεντρικές τιμές.

Όταν κατασκευάζουμε μια κατανομή συχνοτήτων, είναι επιθυμητό το εύρος των ταξικών διαστημάτων να είναι ίδιο για όλες τις τάξεις. Αυτό επιτρέπει τη σύγκριση των συχνοτήτων που αντιστοιχούν σε δύο διαφορετικά ταξικά διαστήματα και διευκολύνει τους υπολογισμούς των παραμέτρων της κατανομής. Όταν όμως ένα μεγάλο ποσοστό των παρατηρήσεων βρίσκεται κατανεμημένο σε ορισμένο διάστημα τιμών, το οποίο αποτελεί ένα μικρό μέρος του εύρους των παρατηρήσεων, τότε είμαστε υποχρεωμένοι να χρησιμοποιήσουμε άνισα ταξικά διαστήματα.

Πίνακας 2.8 Απόλυτες (n) και σχετικές (f) συχνότητες ανά ηλικία

ΗΛΙΚΙΑ	n	f
[0, 6)	166	0,075
[6, 12)	146	0,066
[12 ,18)	178	0,080
[18, 24)	229	0,103
[24, 30)	229	0,103
[30, 36)	221	0,100
[36, 42)	217	0,098
[42, 48)	213	0,096
[48, 54)	213	0,096
[54, 60)	139	0,063
[60, 66)	115	0,052
[66, 72)	66	0,030
[72, 78)	52	0,023
[78, 84)	23	0,010
[84, 90)	3	0,001
[90, 96)	4	0,002
Σύνολο	2214	1,000

Πηγή: Μελέτη Διερεύνησης Κοινωνικών Αναγκών στους Δήμους του Νομού Θεσσαλονίκης (2003)