Μέτρα μεταβλητότητας

Αν τα μέτρα θέσης του προηγούμενου κεφαλαίου μας περιγράφουν γύρω από πια τιμή συγκεντρώνονται οι μετρήσεις (τιμές) μίας μεταβλητής, τα μέτρα διασποράς μας δείχνουν πόσο οι μετρήσεις αυτές απομακρύνονται (διασπείρονται) γύρω από τη θέση (π.χ. τη μέση τιμή). Όπως και στην περίπτωση των μέτρων θέσης, δεν υπάρχει ένας μοναδικός τρόπος για να μετρήσουμε τη διασπορά που να είναι ο ιδανικός σε όλες τις περιπτώσεις. Εδώ θα δούμε μέρικα από τα πιο συχνά μέτρα διασποράς όπως είναι το εύρος, το ενδοτεταρτημοριακό εύρος και η διασπορά/τυπική απόκλιση.

Εύρος

Το εύρος είναι το πιο απλό από τα μέτρα διασποράς και υπολογίζεται ως η διαφορά μεταξύ της μέγιστης και της ελάχιστης τιμής μίας μεταβλητής. Συνήθως συμβολίζουμε το εύρος με το γράμμα R, από την αγγλική λέξη *range* και το υπολογίζουμε από:

$$R = max(X) - min(X)$$

Το έυρος, αν και είναι πάρα πολύ εύκολο να υπολογιστεί, δυστυχώς είναι πολύ ευαίσθητο σε ακραίες τιμές. Η ύπαρξη ακόμη και μίας ακραίας τιμής μπορεί να δώσει λανθασμένη εντύπωση της πραγματικής μεταβλητότητας.

Ενδοτεταρτημοριακό εύρος

Η ιδέα είναι ίδια με αυτή του εύρους, ωστόσο, υπολογίζουμε τη διαφορά μεταξύ του πρώτου και του τρίτου τερτημορίου των τιμών μιας μεταβλητής. Το ενδοτεταρτημοριακό εύρος το συμβολίζουμε συνήθως με IQR από την αγγλική ορολογία *interquartile range* και το υπολογίζουμε από:

$$IQR = Q3 - Q1 = P_{75} - P_{25}$$

Το ενδοτεταρτημοριακό εύρος δε μας δίνει πόσο απέχουν μεταξύ τους η ελάχιστη και η μέγιστη τιμή των μετρήσεων μιας μεταβλητής αλλά την απόσταση μεταξύ δύο τιμών ανάμεσα στις οποίες βρίσκεται το 50% των (μεσαίων) μετρήσεων μιας μεταβλητής. Το ενδοτεταρτημοριακό εύρος έχει μεγαλύτερη αντοχή στις ακραίες τιμές, αφού έχει μεταφέρει το ενδιαφέρον στις μεσαίες τιμές μιας μεταβλητής, αποκλείοντας το 25% των μεγαλύτερων και των μικρότερων τιμών.

Διασπορά και τυπική απόκλιση

Μέχρι τώρα ποσοτικοποιούμε τη μεταβλητότητα των μετρήσεων μίας μεταβλητής μετρώντας αποστάσεις μεταξύ των ακραίων τιμών της, είτε αυτές είναι το μέγιστο και το ελάχιστο ή κάποια τεταρτημόρια. Ένας άλλος τρόπος είναι να υπολογίσουμε με κάποιον τρόπο πόσο απομακρύνονται οι τιμές μίας μεταβλητής από το κέντρο (τη θέση, π.χ. τη μέση τιμή). Στη στατιστική, αποκαλούμε τις αποστάσεις των μετρήσεων από τη μέση τιμή **αποκλίσεις**. Με συμβολισμούς, απόκλιση ονομάζεται η ποσότητα

$(x_i - \bar{x})$, όπου x_i είναι οι τιμές της μεταβλητής X. Αν το μέγεθος του δείγματός μας είναι n, τότε μπορούμε να υπολογίσουμε n αποκλίσεις από τη μέση τιμή $\bar{x}$.

Για τον υπολογισμό της διασποράς της μεταβλητής X γύρω από τη μέση τιμή της πρέπει με κάποιον τρόπο να υπολογίσουμε το μέσο όρο των αποκλίσεων όλων των τιμών της X. Δυστυχώς, το αν πάμε να υπολογίσουμε τον μέσο όρο των αποκλίσεων θα δούμε ότι αυτός θα είναι πάνοτε 0:

$$\begin{aligned}\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n} &= \frac{(x_1 - \bar{x}) + (x_2 - \bar{x}) + ... + (x_n - \bar{x})}{n} \\ &= \frac{(x_1 + x_2 + ... + x_n) - n\bar{x}}{n} \\ &= \frac{x_1 + x_2 + ... + x_n}{n} - \frac{n\bar{x}}{n} \\ &= \bar{x} - \bar{x} = 0\end{aligned}$$

Μπορείτε να το επιβεβαιώσετε αυτό με αν προσπαθήσετε να υπολογίσετε τον μέσο όρο των αποκλίσεων σε ένα μικρό δείγμα, π.χ. $X = [1, 2, 4, 5]$.

Το πρόβλημα προκύπτει από το γεγονός ότι οι αποκλίσεις παίρνουν και αρνητικές και θετικές τιμές, αναλόγως κάθε φορά αν η τιμή της X (δηλαδή το x_i) είναι μικρότερο ή μεγαλύτερο από τη μέση τιμή. Όταν το x_i είναι μεγαλύτερο από τη μέση τιμή, τότε η απόκλιση $(x_i - \bar{x}$ θα είναι θετική. Στην αντίθετη περίπτωση, θα είναι αρνητική.

Για τον υπολογισμό της διασποράς δε μας ενδιαφέρει αν η τιμές της X είναι πάνω ή κάτω από τη μέση τιμή. Το μόνο που μας ενδιαφέρει είναι η απόσταση αυτή καθεαυτή. Άρα θέλουμε να κάνουμε μια μετατροπή στην τιμή της απόκλισης, ώστε το αποτέλεσμα να είναι πάντα θετικό. Μία τέτοια μετατροπή, ιδιαίτερα χρήσιμη στη στατιστική, είναι να υψώσουμε όλες τις αποκλίσεις στο τετράγωνο, δηλαδή να υπολογίσουμε το $(x_i - \bar{x})^2$ αντί του $(x_i - \bar{x})$. Τώρα μπορούμε να υπολογίσουμε τον μέσο όρο των τετραγωνικών αποκλίσεων:

$$s_n^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n}$$

Η ποσότητα s_n^2 ονομάζεται **διακύμανση** της μεταβλητής X. Επομένως, με απλά λόγια, διακύμανση είναι ο μέσος όρος των τετραγωνικών αποκλίσεων των τιμών της μεταβλητής X από τη μέση τιμή.

Δυστυχώς, αν και η διακύμανση είναι ένα αξιόπιστο μέτρο μεταβλητότητας, έχει ένα σημαντικό μειονέκτημα. Δεν εκφράζεται με τις ίδιες μονάδες όπως η μεταβλητή X. Ενώ εμάς μας ενδιαφέρει η μέση απόκλιση των τιμών της X από τη μέση τιμή της, η διακύμανση υψώνει αυτές τις αποκλίσεις στο τετράγωνο. Επομένως, η διακύμανση δεν είναι άμεσα ερμηνεύσιμη. Δεν μπορώ να τη χρησιμοποιήσω απευθείας για να

προβλέψω πόση περιμένω να είναι η απόσταση μίας νέας μέτρησης από τη μέση τιμή. Αντίθετα, η διακύμανση μου λέει πόσο περιμένω να είναι το τετράγωνο αυτής της απόστασης. Επομένως, αν πάρουμε την τετραγωνική ρίζα της διακύμανσης (s_n^2), θα μπορέσουμε να βρούμε αυτό που ζητάμε.

Η ποσότητα s_n,

$$s_n = \sqrt{s_n^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

ονομάζεται **τυπική απόκλιση**. Η τυπική απόκλιση s_n, μετράται στην ίδια μονάδα μέτρησης με την μεταβλητή X και επομένως μπορεί να χρησιμοποιηθεί για να περιγράψει το μέγεθος των αποστάσεων που αναμένουμε για τις τιμές της X από τη μέση τιμή της.