Introduction to R for Social Sciences

Περιγραφική στατιστική

Αναστάσιος Εμβαλωτής & Αικατερίνη Σαργιώτη

Μέτρα θέσης (μέτρα κεντρικής τάσης)

- Προσδιορίζουν ένα σημείο γύρω από το οποίο έχουν την τάση να συγκεντρώνονται τα δεδομένα
- Αναφέρονται σε συνεχή δεδομένα
- Κυριότερα μέτρα θέσης:
  1. Μέση τιμή (mean)
  2. Διάμεσος (median)
  3. Τεταρτημόρια (quartiles)
  4. Ποσοστιαία σημεία (percentiles)
  5. Επικρατούσα τιμή (mode)

Εφαρμογή στην R

- Για να υπολογιστούν τα μέτρα θέσης στην R, θα πρέπει αρχικά να γίνει η εισαγωγή της βάσης δεδομένων
  - *variables_1 = c("ST004D01T", "ST123Q02NA", "Math_Score" , "Science_Score")*
  - *dataset_1 = PISA2015Lab_1[variables_1]*

Μέση τιμή (mean)

- mean = $\frac{\quad\quad\quad\quad\quad\quad\quad\quad}{}$

- Υπολογίζουμε τη μέση τιμή συναρτήσει του αρχείου που έχουμε εισάγει και της μεταβλητής που θέλουμε να υπολογίσουμε, χρησιμοποιώντας ως διαχωριστή το $

mean(dataset_1$Math_Score)

Διάμεσος (median)

- Περιττός αριθμός παρατηρήσεων → Η διάμεσος είναι η μεσαία παρατήρηση
  - Υπολογισμός της θέσης της διαμέσου:
  - $= \frac{}{}$
- Άρτιος αριθμός παρατηρήσεων → Η διάμεσος είναι η μέση τιμή των δύο μεσαίων παρατηρήσεων
  - Υπολογισμός των θέσεων των δύο μεσαίων παρατηρήσεων:
  - $1 = -$ και $2 = \frac{}{}$
- Υπολογίζουμε τη διάμεσο συναρτήσει του αρχείου που έχουμε εισάγει και της μεταβλητής που θέλουμε να υπολογίσουμε, χρησιμοποιώντας ως διαχωριστή το $

median(dataset_1$Math_Score)

Τεταρτημόρια (quantiles)

- Υπολογίζουμε τα τεταρτημόρια συναρτήσει του αρχείου που έχουμε εισάγει και της μεταβλητής που θέλουμε να υπολογίσουμε, χρησιμοποιώντας ως διαχωριστή το $

quantile(dataset_1$Math_Score)

Ποσοστιαία σημεία (percentiles)

- Υπολογίζουμε τα ποσοστιαία σημεία συναρτήσει του αρχείου που έχουμε εισάγει και της μεταβλητής που θέλουμε να υπολογίσουμε, χρησιμοποιώντας ως διαχωριστή το $, αλλά και εισάγοντας στην εντολή τα σημεία που θέλουμε

quantile(dataset_1$Math_Score, c(.32,.57,.88))

Μέτρα διασποράς

- Προσδιορίζουν το πόσο «διεσπαρμένα» είναι τα δεδομένα
- Αναφέρονται σε συνεχή δεδομένα
- Κυριότερα μέτρα διασποράς:
    1. Διακύμανση ή διασπορά (variance)
    2. Τυπική απόκλιση (standard deviation)
    3. Εύρος (range)
    4. Ενδοτεταρτημοριακό εύρος (interquartile range)

Εφαρμογή στην R

- Για να υπολογιστούν τα μέτρα θέσης στην R, θα πρέπει αρχικά να γίνει η εισαγωγή της βάσης δεδομένων
  - *variables_1 = c(“Math_Score”)*
  - *dataset_1 = PISA2015Lab_1[variables_1]*

Διακύμανση ή διασπορά (variance)

- Υπολογίζουμε τη διακύμανση/διασπορά συναρτήσει του αρχείου που έχουμε εισάγει και της μεταβλητής που θέλουμε να υπολογίσουμε, χρησιμοποιώντας ως διαχωριστή το $

var(dataset_1$Math_Score)

Τυπική απόκλιση (standard deviation)

- $s = \sqrt{\ }$
- Υπολογίζουμε την τυπική απόκλιση συναρτήσει του αρχείου που έχουμε εισάγει και της μεταβλητής που θέλουμε να υπολογίσουμε, χρησιμοποιώντας ως διαχωριστή το $

sd(dataset_1$Math_Score)

Εύρος (range)

- range = max - min
- Μέγιστη τιμή (συναρτήσει του αρχείου που έχουμε εισάγει και της μεταβλητής που θέλουμε να υπολογίσουμε, χρησιμοποιώντας ως διαχωριστή το $)

max(dataset_1$Math_Score)

- Ελάχιστη τιμή (συναρτήσει του αρχείου που έχουμε εισάγει και της μεταβλητής που θέλουμε να υπολογίσουμε, χρησιμοποιώντας ως διαχωριστή το $)

min(dataset_1$Math_Score)

- Εύρος (συναρτήσει του αρχείου που έχουμε εισάγει, της μεταβλητής που θέλουμε να υπολογίσουμε, της max και min, χρησιμοποιώντας ως διαχωριστή το $)

max(dataset_1$Math_Score)- min(dataset_1$Math_Score)

Ενδοτεταρτημοριακό εύρος (interquartile range)

- IQR = 3ο τεταρτημόριο – 1ο τεταρτημόριο
- Υπολογίζουμε το ενδοτεταρτημοριακό εύρος συναρτήσει του αρχείου που έχουμε εισάγει και της μεταβλητής που θέλουμε να υπολογίσουμε, χρησιμοποιώντας ως διαχωριστή το $

IQR(dataset_1$Math_Score)

Τυπικό σφάλμα (standard error)

- SE = $\frac{\cdot}{\sqrt{}}$
- Η τυπική απόκλιση υπολογίζεται με την εντολή *sd()*
- Το μέγεθος του δείγματος υπολογίζεται με την εντολή *length()*

n = length(dataset_1$Math_Score)

SE = sd(dataset_1$Math_Score) / sqrt(dataset_1$Math_Score)

Άλλα μέτρα περιγραφικής στατιστικής

1. Συντελεστές ασυμμετρίας (skewness coefficients)
2. Συντελεστής μεταβλητότητας (CV-coefficient of variation)
3. Συντελεστής κύρτωσης (kurtosis coefficient)

Εφαρμογή στην R

- Για να υπολογιστούν τα μέτρα θέσης στην R, θα πρέπει αρχικά να γίνει η εισαγωγή της βάσης δεδομένων
  - *variables_1 = c(“Math_Score”)*
  - *dataset_1 = PISA2015Lab_1[variables_1]*

Συντελεστής ασυμμετρίας (skewness coefficients)

- Πρέπει, αρχικά, να εγκαταστήσουμε και να φορτώσουμε τη βιβλιοθήκη που έχει μέσα τη συνάρτηση της ασυμμετρίας, συγκεκριμένα το πακέτο **“e1071”**

install.pacjages(“e1071”)

library(e1071)

- Υπολογίζουμε την ασυμμετρία συναρτήσει του αρχείου που έχουμε εισάγει και της μεταβλητής που θέλουμε να υπολογίσουμε, χρησιμοποιώντας ως διαχωριστή το $

skewness(dataset_1$Math_Score)

Συντελεστής μεταβλητότητας (CV-coefficient of variation)

- CV = $\frac{\cdot}{\cdot\cdot} \cdot 100\%$
- Υπολογίζουμε τον συντελεστή μεταβλητότητας συναρτήσει του αρχείου που έχουμε εισάγει και της μεταβλητής που θέλουμε να υπολογίσουμε, χρησιμοποιώντας ως διαχωριστή το $

sd(dataset_1$Math_Score)/mean(dataset_1$Math_Score)

Συντελεστής κύρτωσης (kurtosis coefficient)

- Πρέπει, αρχικά, να εγκαταστήσουμε και να φορτώσουμε τη βιβλιοθήκη που έχει μέσα τη συνάρτηση της κύρτωσης, συγκεκριμένα το πακέτο **“e1071”**

install.pacjages(“e1071”)

library(e1071)

- Υπολογίζουμε την κύρτωση συναρτήσει του αρχείου που έχουμε εισάγει και της μεταβλητής που θέλουμε να υπολογίσουμε, χρησιμοποιώντας ως διαχωριστή το $

kurtosis(dataset_1$Math_Score)

Εντολή summary()

- Μας δίνει ορισμένα μέτρα θέσης και διασποράς

```
summary(dataset_1$Math_Score)
```