
Introduction to R for Social Sciences

Πίνακες και γραφήματα περιγραφικής στατιστικής

Αναστάσιος Εμβαλωτής & Αικατερίνη Σαργιώτη

Πίνακες

1. Πίνακες συχνοτήτων
 - Μπορεί να χρησιμοποιηθεί για συνεχείς ή διακριτές μεταβλητές
2. Πίνακες συνάφειας
 - Χρησιμοποιείται για δύο ποιοτικές μεταβλητές

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]`

Πίνακες συχνοτήτων (1)

- Δημιουργούμε πίνακα με τις μεταβλητές που θέλουμε (διακριτές/συνεχείς μεταβλητές) με τις εντολές:
 - `table ()`, για πίνακες απόλυτων συχνοτήτων, και
 - `prop.table (table())`, για πίνακες σχετικών συχνοτήτων (ποσοστά)

Πίνακες συχνοτήτων (2)

Διακριτή μεταβλητή

(εάν οι γονείς υποστηρίζουν τις εκπαιδευτικές προσπάθειες και τα επιτεύγματα των παιδιών τους)

- Πίνακας **απόλυτων** συχνοτήτων

`table(dataset_1$ST123Q02NA)`

- Πίνακας **σχετικών** συχνοτήτων

`prop.table(table(dataset_1$ST123Q02NA))`

Συνεχής μεταβλητή

(επίδοση στα μαθηματικά)

- Πίνακας **απόλυτων** συχνοτήτων

`table(dataset_1$Math_Score)`

- Πίνακας **σχετικών** συχνοτήτων

`prop.table(table(dataset_1$Math_Score))`

Πίνακες Συνάφειας (1)

- Δημιουργούμε πίνακα με τις μεταβλητές που θέλουμε (διακριτές μεταβλητές) με τις εντολές:
 - `table(x,y)`, απόλυτες συχνότητες, και
 - `prop.table(table(x,y))`, σχετικές συχνότητες (ποσοστά)

Πίνακες Συνάφειας (2)

Διακριτές μεταβλητές

(φύλο & εάν οι γονείς υποστηρίζουν τις εκπαιδευτικές προσπάθειες και τα επιτεύγματα των παιδιών τους)

- Πίνακας συνάφειας **απόλυτων** συχνοτήτων
 - `Table_1 = table(dataset_1$ST004D01T, dataset_1$ST123Q02NA)`
 - `Table_1`
- Εμφάνιση απόλυτων συχνοτήτων της πρώτης μεταβλητής του πίνακα συνάφειας (φύλο)
 - `margin.table(Table1, 1)`
- Εμφάνιση απόλυτων συχνοτήτων της δεύτερης μεταβλητής του πίνακα συνάφειας (εάν οι γονείς υποστηρίζουν τις εκπαιδευτικές προσπάθειες και τα επιτεύγματα των παιδιών τους)
 - `margin.table(Table1, 2)`
- Πίνακας συνάφειας **σχετικών** συχνοτήτων
 - `Table_2 = prop.table(Table_1)`
 - `Table_2`

Διαγράμματα περιγραφικής στατιστικής

1. Ιστόγραμμα (Histogram)
2. Θηκόγραμμα (box-plot)
3. Κυκλικό διάγραμμα (pie chart)
4. Διάγραμμα σκεδασμού (scatter plot)
5. Ραβδόγραμμα (bar chart)

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]`

Ιστόγραμμα (Histogram) (1)

- Συνεχείς μεταβλητές
- Δημιουργούμε το ιστόγραμμα συναρτήσει του αρχείου που έχουμε εισάγει και της μεταβλητής που θέλουμε να υπολογίσουμε, χρησιμοποιώντας ως διαχωριστή το `$`, με την εντολή `hist()`
 - `hist(dataset_1$Math_Score)`
 - `hist(dataset_1$Math_Score, right = F)`
- Αλλαγή ιδιοτήτων ιστογράμματος
 - `hist(dataset_1$Math_Score, right=F, col="purple", main="Performance in Maths", xlab="Students' Performance")`
- Σχεδιασμός καμπύλης κανονικής κατανομής με την εντολή `lines(density(x))`, x η συνεχής μεταβλητή
 - `hist(dataset_1$Math_Score, right=F, freq=F)`
 - `lines(density(dataset_1$Math_Score), col="blue")`

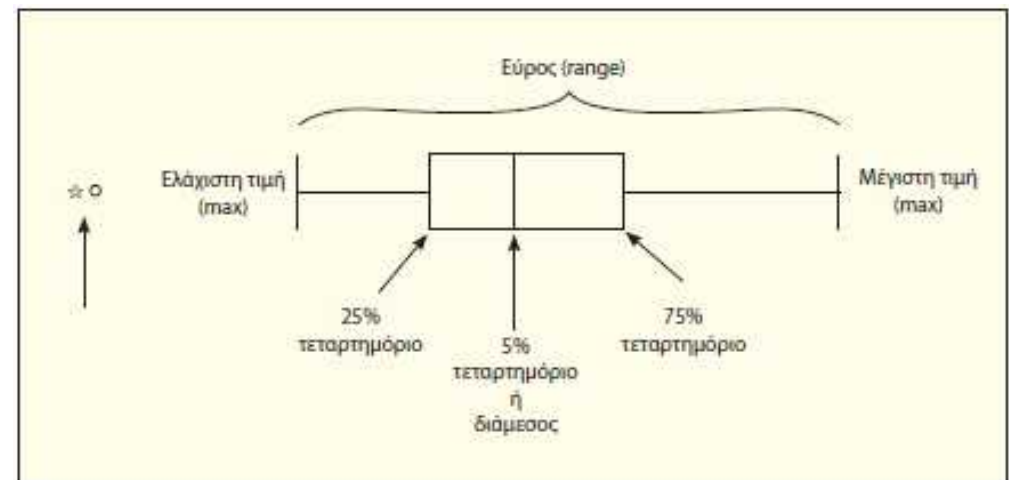
Ιστόγραμμα (Histogram) (2)

Δημιουργία ενός ιστογράμματος μιας συνεχούς μεταβλητής με βάση μια διακριτή
(διακριτή μεταβλητή: Φύλο = 1 → Γυναίκες)

- Για να δημιουργήσουμε το ιστόγραμμα μιας συνεχούς μεταβλητής με βάση μια διακριτή χρησιμοποιούμε την εντολή `hist()`, τις δύο μεταβλητές και στην διακριτή ορίζουμε την τιμή που θέλουμε να πάρει
 - `hist(dataset_1$Math_Score[dataset_1$ST004D01T==1])`
- Για αποθήκευση του ιστογράμματος επιλέγουμε το περιβάλλον του ιστογράμματος και στη συνέχεια
 - `File` → `Save as` → τον τύπο του αρχείου που θέλουμε να γίνει η αποθήκευση του ιστογράμματος

Θηκόγραμμα (box-plot) (1)

- Συνεχείς μεταβλητές
- Δημιουργούμε το θηκόγραμμα συναρτήσει του αρχείου που έχουμε εισάγει και της μεταβλητής που θέλουμε να υπολογίσουμε, χρησιμοποιώντας ως διαχωριστή το \$, με την εντολή `boxplot()`
 - `boxplot(dataset_1$Math_Score)`



Θηκόγραμμα (box-plot) (2)

- Αλλαγή ιδιοτήτων Θηκογράμματος

- `boxplot(dataset_1$Math_Score, col="purple", main="Performance in Maths")`

Δημιουργία ενός Θηκογράμματος μιας συνεχούς μεταβλητής με βάση μια διακριτή
(διακριτή μεταβλητή: Φύλο = 1 → Γυναίκες)

- Για να δημιουργήσουμε το θηκόγραμμα μιας συνεχούς μεταβλητής με βάση μια διακριτή χρησιμοποιούμε την εντολή `boxplot(x[y==a])`, τις δύο μεταβλητές και στην διακριτή ορίζουμε την τιμή που θέλουμε να πάρει
 - `boxplot(dataset_1$Math_Score[dataset_1$ST004D01T==1])`
- Για αποθήκευση του θηκογράμματος επιλέγουμε το περιβάλλον του θηκογράμματος και στη συνέχεια
 - `File` → `Save as` → τον τύπο του αρχείου που θέλουμε να γίνει η αποθήκευση του θηκογράμματος

Κυκλικό διάγραμμα (pie chart)

- Διακριτές μεταβλητές

Κυκλικό διάγραμμα απόλυτων συχνοτήτων

- Αρχικά, δημιουργείται πίνακας απόλυτων συχνοτήτων
 - `table(dataset_1$ST004D01T)`
- Κατασκευή διαγράμματος με την εντολή `pie()`
 - `pie(table(dataset_1$ST004D01T))`

Κυκλικό διάγραμμα σχετικών συχνοτήτων

- Αρχικά, δημιουργείται πίνακας σχετικών συχνοτήτων
 - `prop.table(dataset_1$ST004D01T)`
- Κατασκευή διαγράμματος με την εντολή `pie()`
 - `pie(prop.table(dataset_1$ST004D01T))`

Για αποθήκευση του κυκλικού διαγράμματος επιλέγουμε το περιβάλλον του διαγράμματος και στη συνέχεια

- `File` → `Save as` → τον τύπο του αρχείου που θέλουμε να γίνει η αποθήκευση του κυκλικού διαγράμματος

Διάγραμμα σκεδασμού (scatter plot) (1)

- Δύο συνεχείς μεταβλητές
- Δημιουργούμε το διάγραμμα σκεδασμού συναρτήσει του αρχείου που έχουμε εισάγει και των μεταβλητών που θέλουμε να υπολογίσουμε, χρησιμοποιώντας ως διαχωριστή το \$, με την εντολή `plot()`
 - `plot(dataset_1$Math_Score, dataset_1$Science_Score, xlab="Performance in Maths", ylab="Performance in Maths")`
- Αλλαγή ορίων διαγράμματος σκεδασμού με τα ορίσματα `xlim=c()` και `ylim=c()`
 - `plot(dataset_1$Math_Score, dataset_1$Science_Score, xlab="Performance in Maths", ylab="Performance in Maths", xlim=c(300,700), ylim=c(400,600))`

Διάγραμμα σκεδασμού (scatter plot) (2)

- Αλλαγή του μεγέθους των σημείων (dots), προσθέτοντας στην εντολή το όρισμα `cex=`
 - `plot(dataset_1$Math_Score, dataset_1$Science_Score, xlab="Performance in Maths", ylab="Performance in Maths", xlim=c(300,700), ylim=c(400,600), cex=0.5)`
- Αλλαγή του χρώματος των σημείων (dots), προσθέτοντας στην εντολή το όρισμα `col = "colors"`
 - `plot(plot(dataset_1$Math_Score, dataset_1$Science_Score, xlab="Performance in Maths", ylab="Performance in Maths", xlim=c(300,700), ylim=c(400,600), col="purple")`
- Κατασκευή γραμμής παλινδρόμησης με την εντολή `abline(x, y, col = "colors")`
 - `abline(lm(dataset_1$Math_Score ~ dataset_1$Science_Score), col="blue")`
- Για αποθήκευση του διαγράμματος σκεδασμού επιλέγουμε το περιβάλλον του διαγράμματος και στη συνέχεια
 - `File` → `Save as` → τον τύπο του αρχείου που θέλουμε να γίνει η αποθήκευση του διαγράμματος σκεδασμού

Ραβδόγραμμα (bar chart) (1)

- Διακριτές μεταβλητές

Απλό ραβδόγραμμα

- Αρχικά κατασκευάζεται ο πίνακας απόλυτων ή σχετικών συχνοτήτων
- Δημιουργούμε το ραβδόγραμμα συναρτήσεως του αρχείου που έχουμε εισάγει και των μεταβλητών που θέλουμε να υπολογίσουμε, χρησιμοποιώντας ως διαχωριστή το \$, με την εντολή `barplot(table(x))` και `barplot(prop.table(x))`
 - `barplot(table(dataset_1$ST004D01T))`
 - `barplot(prop.table(table(dataset_1$ST004D01T)))`
- Αλλαγή ιδιοτήτων
 - `barplot(table(dataset_1$ST004D01T), col="purple", main="Gender", xlab="Gender", ylab="Count")`
→ ίδιο χρώμα σε όλες τις ράβδους
 - `barplot(table(dataset_1$ST004D01T), col=c("purple", "blue"), main="Gender", xlab="Gender", ylab="Count")` → διαφορετικό χρώμα στις ράβδους

Ραβδόγραμμα (bar chart) (2)

Απλό οριζόντιο ραβδόγραμμα

- Αρχικά κατασκευάζεται ο πίνακας απόλυτων ή σχετικών συχνοτήτων
- Δημιουργούμε το ραβδόγραμμα συναρτήσει του αρχείου που έχουμε εισάγει και των μεταβλητών που θέλουμε να υπολογίσουμε, χρησιμοποιώντας ως διαχωριστή το \$, με την εντολή `barplot(table(x))` και `barplot(prop.table(x))`
- Προσθήκη του ορίσματος `horiz = T`
 - `barplot(table(dataset_1$ST004D01T), horiz=T, names.arg=c("Girl", "Boy"))`
 - `barplot(prop.table(table(dataset_1$ST004D01T)), horiz=T, names.arg=c("Girl", "Boy"))`
- Αλλαγή ιδιοτήτων
 - `barplot(table(dataset_1$ST004D01T), horiz=T, names.arg=c("Girl", "Boy"), col="purple", main="Gender", xlab="Gender", ylab="Count")` → ίδιο χρώμα σε όλες τις ράβδους
 - `barplot(table(dataset_1$ST004D01T), horiz=T, names.arg=c("Girl", "Boy"), col=c("purple", "blue"), main="Gender", xlab="Gender", ylab="Count")` → διαφορετικό χρώμα στις ράβδους

Ραβδόγραμμα (bar chart) (3)

Ομαδοποιημένο ραβδόγραμμα

- Δύο διακριτές μεταβλητές (φύλο & εάν οι γονείς υποστηρίζουν τις εκπαιδευτικές προσπάθειες και τα επιτεύγματα των παιδιών τους)
- Αρχικά, δημιουργείται ο πίνακας συνάφειας των δύο διακριτών μεταβλητών
 - `bartable_1 = table(dataset_1$ST004D01T, dataset_1$ST123Q02NA)`
- Δημιουργούμε το ραβδόγραμμα συναρτήσεως του αρχείου που έχουμε εισάγει και των μεταβλητών που θέλουμε να υπολογίσουμε, χρησιμοποιώντας ως διαχωριστή το \$, με την εντολή `barplot(bartable())`
 - `barplot(bartable_1, beside=T, legend=levels(unique(dataset_1$ST004D01T)))`

Ραβδόγραμμα (bar chart) (4)

Οριζόντιο ομαδοποιημένο ραβδόγραμμα

- Δύο διακριτές μεταβλητές (φύλο & εάν οι γονείς υποστηρίζουν τις εκπαιδευτικές προσπάθειες και τα επιτεύγματα των παιδιών τους)
- Αρχικά, δημιουργείται ο πίνακας συνάφειας των δύο διακριτών μεταβλητών
 - `bartable_2 = table(dataset_1$ST004D01T, dataset_1$ST123Q02NA)`
- Δημιουργούμε το ραβδόγραμμα συναρτήσεως του αρχείου που έχουμε εισάγει και των μεταβλητών που θέλουμε να υπολογίσουμε, χρησιμοποιώντας ως διαχωριστή το \$, με την εντολή `barplot(bartable())`
 - `barplot(bartable_2, beside=T, horiz = T, legend=levels(unique(dataset_1$ST004D01T)))`

Ραβδόγραμμα (bar chart) (5)

- Αλλαγή ονομασίας ράβδων
 - Προσθήκη ορίσματος `names.arg = c("Name_value_1", "Name_value_2", ..., "Name_value_n")`
- Αλλαγή χρώματος ράβδων
 - Προσθήκη ορίσματος `col = c("color")` (ίδιο χρώμα σε όλες τις ράβδους) ή `col = c("color_1", "color_2")` (διαφορετικό χρώμα στις ράβδους)
- Αλλαγή ονόματος ραβδογράμματος
 - Προσθήκη ορίσματος `main = "Title"`
- Αλλαγή ονόματος αξόνων
 - Προσθήκη ορίσματος `xlab = "Title_x"` → αλλαγή ονόματος οριζόντιου άξονα
 - Προσθήκη ορίσματος `ylab = "Title_y"` → αλλαγή ονόματος κάθετου άξονα

Ραβδόγραμμα (bar chart) (6)

Ομαδοποιημένο ραβδόγραμμα σχετικών συχνοτήτων

- Δύο διακριτές μεταβλητές (φύλο & εάν οι γονείς υποστηρίζουν τις εκπαιδευτικές προσπάθειες και τα επιτεύγματα των παιδιών τους)
- Αρχικά, δημιουργείται πίνακας συνάφειας σχετικών συχνοτήτων
 - `prop.table(dataset_1$ST004D01T, dataset_1$ST123Q02NA)`
- Δημιουργούμε το ραβδόγραμμα σχετικών συχνοτήτων συναρτήσει του αρχείου που έχουμε εισάγει και των μεταβλητών που θέλουμε να υπολογίσουμε, χρησιμοποιώντας ως διαχωριστή το \$, με την εντολή `barplot(bartable())`
 - `barplot(prop.table(dataset_1$ST004D01T, dataset_1$ST123Q02NA), beside=T, legend=levels(unique(dataset_1$ST004D01T)))`

Προσθήκη υπομνήματος στο ραβδόγραμμα

- `barplot(legend("topleft", c("Girls", "Boys"), fill=c("purple", "blue")))`
- Για αποθήκευση του ραβδογράμματος επιλέγουμε το περιβάλλον του διαγράμματος και στη συνέχεια
 - `File` → `Save as` → τον τύπο του αρχείου που θέλουμε να γίνει η αποθήκευση του ραβδογράμματος