Οδηγίες χρήσης του R, μέρος 4^{ο}

(συμπλήρωμα για την εργασία)

Για την ολοκλήρωση της εργασίας, αφού συλλέξετε τα δεδομένα με βάση την προκαθορισμένη διαδικασία μέτρησης, θα πρέπει να εκτελέσετε τις εξής ενέργειες: καταχώριση των δεδομένων, γραφικό και αριθμητικό έλεγχο των δεδομένων, ανάλυση με το κατάλληλο στατιστικό κριτήριο.

Έχοντας περιορίσει τα θέματα εργασίας σε απλή σχέση μεταξύ δύο μεταβλητών, και λαμβάνοντας υπόψη ότι κάθε μεταβλητή μπορεί να είναι ποιοτική (factor) ή ποσοτική (numeric), έχουμε τρεις περιπτώσεις ανάλυσης:

(α) Δύο ποιοτικές μεταβλητές, οπότε κάνουμε σύγκριση συχνοτήτων ή αναλογιών, με το κριτήριο χ^2 (συνάρτηση chisq.test). Σε αυτήν την περίπτωση η μηδενική υπόθεση (H_0) είναι ότι οι αναλογίες συχνότητων της εξαρτημένης μεταβλητής είναι ίσες, δηλαδή $\chi^2 = 0$.

(β) Μία ποιοτική (ανεξάρτητη) και μία ποσοτική (εξαρτημένη) μεταβλητή, οπότε κάνουμε σύγκριση μέσων όρων της εξαρτημένης μεταβλητής ανάμεσα στις κατηγορίες της ανεξάρτητης, με το κριτήριο *t* (συνάρτηση t.test). Σε αυτήν την περίπτωση η μηδενική υπόθεση (H_0) είναι ότι οι μέσοι όροι της εξαρτημένης μεταβλητής είναι ίσοι, δηλαδή $t = 0$.

(γ) Δύο ποσοτικές μεταβλητές, οπότε κάνουμε έλεγχο συσχέτισης με το κριτήριο του συντελεστή συσχέτισης r_{xy} (συνάρτηση cor.test). Σε αυτήν την περίπτωση η μηδενική υπόθεση (H_0) είναι ότι οι μεταβλητές είναι ανεξάρτητες, δηλαδή $r_{xy} = 0$.

Α. Καταχώριση δεδομένων

Για την καταχώριση των δεδομένων μας στο R μπορούμε να χρησιμοποιήσουμε διάφορα άλλα προγράμματα, όπως το Microsoft Excel, ή οποιοδήποτε πρόγραμμα μπορεί να χειριστεί και να αποθηκεύσει δεδομένα με μορφή πίνακα (τύπου csv). Στη συνέχεια εισάγουμε τα στοιχεία σε ένα πλαίσιο δεδομένων στο R διαβάζοντας το σχετικό αρχείο με τη συνάρτηση read.csv.

Για τις ανάγκες της εργασίας, επειδή ο όγκος των δεδομένων είναι μικρός, μπορούμε να χρησιμοποιήσουμε το ίδιο το R. Αυτός είναι ο απλούστερος τρόπος.

Πρώτα ορίζουμε το πλαίσιο δεδομένων με τις δύο μεταβλητές μας. Αν, για παράδειγμα, έχουμε μία ποιοτική ανεξάρτητη μεταβλητή f και μία ποσοτική εξαρτημένη μεταβλητή x, ορίζουμε

```
> d<-data.frame(f=factor(),x=numeric())
```

Δώστε αναγνωρίσιμα ονόματα στις μεταβλητές σας—όχι x και f, αλλά πιο περιγραφικούς μονολεκτικούς όρους, π.χ. «Φύλο», «Ηλικία» (μπορείτε να χρησιμοποιήσετε ελληνικά). Έτσι θα φαίνονται πιο ξεκάθαρα τα αποτελέσματα και οι γραφικές παραστάσεις.

Αντίστοιχα, αν έχουμε δύο ποιοτικές μεταβλητές ορίζουμε και τις δύο ως factor, ενώ αν έχουμε δύο ποσοτικές μεταβλητές ορίζουμε και τις δύο ως numeric. Η σειρά των μεταβλητών στο

πλαίσιο δεδομένων είναι ελεύθερη, βολεύει όμως να προηγείται η ανεξάρτητη μεταβλητή, ειδικά αν είναι τύπου factor, διότι αυτό επιτρέπει ορισμένες αυτόματες λειτουργίες.

Στη συνέχεια εισάγουμε τα δεδομένα από τις μετρήσεις μας καλώντας τη συνάρτηση fix:

```
> fix(d)
```

Στο παράθυρο που θα ανοίξει, θα δούμε ένα διαγραμμισμένο πλαίσιο, στο οποίο οι δύο πρώτες στήλες έχουν ως επικεφαλίδες τα ονόματα των μεταβλητών μας. Συμπληρώνουμε τις δύο αυτές στήλες, προσθέτοντας μια σειρά για κάθε μονάδα του δείγματός μας (π.χ. για κάθε άτομο). Σε περίπτωση ποσοτικής μεταβλητής, δίνουμε με αριθμητική μορφή το αποτέλεσμα της μέτρησης. Δακτυλογραφούμε, δηλαδή, τον αριθμό μέσα στο κουτάκι. Σε ποιοτική μεταβλητή, δακτυλογραφούμε την ετικέτα της κατάλληλης κατηγορίας. Για παράδειγμα, σε μέτρηση φύλου μπορούμε να βάζουμε το γράμμα Α στους άντρες και το γράμμα Γ στις γυναίκες. Ή ολόκληρη τη λέξη «άντρας» / «γυναίκα», αν προτιμάμε (ευκολότερα με copy-paste). Σε άλλη περίπτωση μπορούμε να βάζουμε τη λέξη «Ναι» ή «Όχι» (ή το γράμμα Υ / Ν).

Πριν αρχίσουμε την εισαγωγή των δεδομένων θα πρέπει να αποφασίσουμε ακριβώς ποια θα είναι η μορφή και ο τύπος των μεταβλητών μας και το είδος της ανάλυσης που θα γίνει. Σε περίπτωση ιεραρχικής μεταβλητής, θα πρέπει να αποφασίσουμε αν θα τη χειριστούμε ποιοτικά ή ποσοτικά. Π.χ. μια εξαρτημένη ιεραρχική μεταβλητή όπως το έτος σπουδών μπορεί να κωδικοποιηθεί ως ποσοτική, με αριθμούς 1–4 (για τα έτη Α–Δ) και 5 (για τους επί πτυχίω), αν θέλουμε να υπολογίσουμε και να εξετάσουμε μέσους όρους. Μπορεί όμως να κωδικοποιηθεί ως ποιοτική, με γράμματα Α–Δ και Π, αν θέλουμε να συγκρίνουμε αναλογίες. Η ποιοτική κωδικοποίηση είναι πάντα σωστή, μπορεί όμως να μην εξυπηρετεί το σκοπό μας. Η ποσοτική κωδικοποίηση είναι θεωρητικά λανθασμένη, μπορεί όμως να είναι αποδεκτή σε κάποιες περιπτώσεις, ανάλογα με τη μεταβλητή και την επιθυμητή ανάλυση.

Σε περίπτωση κατηγορικής μεταβλητής πρέπει να είναι σαφείς οι κατηγορίες που εισάγονται στο πλαίσιο δεδομένων. Θα πρέπει να είμαστε πολύ προσεκτικοί στην πληκτρολόγηση, διότι αν μια φορά π.χ. κάνουμε λάθος και γράψουμε «Να» αντί για «Ναι» (ή χρησιμοποιήσουμε λατινικό Ν αντί για ελληνικό) θα μετρήσει ως νέα ετικέτα, δηλαδή ξεχωριστή κατηγορία.

Όταν ολοκληρώσουμε την καταχώριση των δεδομένων, κλείνουμε το παράθυρο εισαγωγής με κλικ πάνω στο ✕ (στην πάνω δεξιά γωνία), οπότε η συνάρτηση fix διατηρεί στη μεταβλητή που ορίσαμε ό,τι δεδομένα πληκτρολογήσαμε στα κουτάκια. Αποθηκεύουμε αμέσως το πλαίσιο δεδομένων στο δίσκο για να μην το χάσουμε, χρησιμοποιώντας τη συνάρτηση write.table:

```
> write.table(d,"ergasia.Rdata")
```

Φυσικά μπορούμε να χρησιμοποιήσουμε όποιο όνομα αρχείου θέλουμε, αντί για ergasia.Rdata, αρκεί να το θυμόμαστε για να το βρούμε αργότερα, όποτε θελήσουμε να φορτώσουμε τα δεδομένα αυτά προς επεξεργασία.

Β. Έλεγχος δεδομένων

Για τον έλεγχο των δεδομένων μας θα χρησιμοποιήσουμε τις γνωστές συναρτήσεις.

```
> str(d)
> summary(d)
```

Από την απάντηση που θα μας δείξει το R θα πρέπει να βεβαιωθούμε ότι δεν έχει γίνει λάθος στην καταχώριση. Για παράδειγμα, οι ελάχιστες και μέγιστες τιμές πρέπει να είναι εύλογες. Το πλήθος και οι ετικέτες των κατηγοριών (για ποιοτικές μεταβλητές) πρέπει να είναι σωστό. Το πλήθος των δεδομένων (αριθμός εγγραφών) πρέπει να είναι σωστό.

Σε περίπτωση που υποψιαστούμε ή διαπιστώσουμε κάποιο λάθος, πρέπει αμέσως να το διορθώσουμε. Η εγκυρότητα της ανάλυσης εξαρτάται απόλυτα από την αξιοπιστία των δεδομένων μας! Χρησιμοποιούμε πάλι τη συνάρτηση fix για να διορθώσουμε ό,τι χρειάζεται. Μετά από κάθε διόρθωση, αποθηκεύουμε το διορθωμένο πλαίσιο δεδομένων (με write.table)!

Αν έχουμε ποσοτικές μεταβλητές, καλό είναι να ελέγξουμε και τις κατανομές των τιμών, με φυλλόγραμμα, ή καλύτερα με ιστόγραμμα (δίνοντας το όνομα της μεταβλητής μας αντί για x):

```
> hist(d$x)
```

Για ένα αδρό έλεγχο κανονικότητας, μπορούμε να χρησιμοποιήσουμε τη συνάρτηση histnorm από το αρχείο pnorm-showz (το οποίο πρέπει να βρίσκεται στον επιλεγμένο φάκελο εργασίας)

```
> source("pnorm-showz.R")
> histnorm(d$x)
```

Αν όλα φαίνονται εντάξει, μπορούμε να εξετάσουμε γραφικά τα δεδομένα μας σε συνδυασμό των δύο μεταβλητών. Για το σκοπό αυτό η συνάρτηση plot μπορεί να «μαντέψει» το είδος των μεταβλητών και, συνήθως, να μας δώσει την κατάλληλη γραφική παράσταση από μόνη της.

```
> plot(d)
```

Γ. Ανάλυση δεδομένων

Έχοντας ελέγξει και επαληθεύσει τα δεδομένα μας, μπορούμε να προχωρήσουμε στην ανάλυση. Για ποσοτικές μεταβλητές, αυτή περιλαμβάνει οπωσδήποτε τον υπολογισμό μέσων όρων και τυπικών αποκλίσεων.

```
> mean(d$x)
> sd(d$x)
```

Αν έχουμε δύο ποσοτικές μεταβλητές στο πλαίσιο δεδομένων, η ανάλυση και των δύο μπορεί να γίνει αυτόματα με μία μόνο κλήση κάθε συνάρτησης:

```
> mean(d)
> sd(d)
```

Εννοείται ότι στην έκθεσή μας εμφανίζουμε το μέσο όρο και την τυπική απόκλιση με τον ίδιο αριθμό δεκαδικών ψηφίων, στρογγυλοποιώντας όπου χρειάζεται ώστε να μην έχουμε πάνω από δύο ψηφία συνολικά στην τυπική απόκλιση.

Αν έχουμε δύο ποιοτικές μεταβλητές τότε πρέπει να υπολογίσουμε τον πίνακα συχνοτήτων:

```
> table(d)
```

Αν για κάποιο λόγο θέλουμε διαφορετική σειρά παρουσίασης των κατηγοριών από εκείνη που μας δίνει το R (που είναι αλφαβητική), τότε μπορούμε να χρησιμοποιήσουμε τον τύπο διατεταγμένης κατηγορικής μεταβλητής (ordered factor). Για παράδειγμα, αν έχουμε τις κατηγορίες “Low”, “Medium” και “High”, μπορούμε να τις αναδιατάξουμε ως εξής:

```
> d$f <- ordered(d$f,levels=c("Low","Medium","High"))
```

Σε περίπτωση δύο ποιοτικών μεταβλητών, η σύγκριση των αναλογιών γίνεται άμεσα:

```
> chisq.test(table(d))
```

Φυσικά, αν υπάρχουν περισσότερες μεταβλητές μέσα στο πλαίσιο δεδομένων θα πρέπει να φροντίσουμε να δώσουμε στη συνάρτηση chisq.test τον πίνακα μόνο των δύο μεταβλητών που μας ενδιαφέρουν, δηλαδή της ανεξάρτητης και της εξαρτημένης, όπως έχουν οριστεί.

```
> chisq.test(table(d[,c(1,2)]))
```

Στη μορφή αυτή, αριστερά από το κόμμα δεν υπάρχει τίποτα, υποδηλώνοντας επιλογή όλων ανεξαιρέτως των σειρών (δηλαδή των μονάδων του δείγματος), ενώ δεξιά υπάρχει η ακολουθία c(1,2), υποδηλώνοντας επιλογή της πρώτης και δεύτερης στήλης (μεταβλητής).

Σε περίπτωση δύο ποσοτικών μεταβλητών, ο έλεγχος συσχέτισης γίνεται άμεσα:

```
> cor.test(d$x,d$y)
```

Εδώ υποθέτουμε ότι οι δύο μεταβλητές στο πλαίσιο δεδομένων ονομάζονται x και y.

Σε περίπτωση κατηγορικής ανεξάρτητης μεταβλητής και ποσοτικής εξαρτημένης μεταβλητής, πριν από τη σύγκριση των μέσων όρων πρέπει να διαχωρίσουμε τις δύο ομάδες βάσει των κατηγοριών της ανεξάρτητης μεταβλητής. Ας υποθέσουμε, για παράδειγμα, ότι η ανεξάρτητη μεταβλητή f τύπου factor περιλαμβάνει τις κατηγορίες “a” και “b”, και η εξαρτημένη μεταβλητή τύπου numeric ονομάζεται x. Τότε σχηματίζουμε τις δύο ομάδες, xa και xb, επιλέγοντας τις κατάλληλες μονάδες από το δείγμα ως εξής:

```
> xa <- d$x[d$f=="a"]
> xb <- d$x[d$f=="b"]
```

Καθεμιά από τις δύο νέες μεταβλητές περιέχει ένα επιλεγμένο υποσύνολο των τιμών της μεταβλητής x, και συγκεκριμένα εκείνο που αντιστοιχεί σε μία κατηγορία της μεταβλητής f.

Στη συνέχεια εκτελούμε τη σύγκριση ανάμεσα στις δύο επιλεγμένες ομάδες με το κριτήριο t:

```
> t.test(xa,xb)
```

Δ. Αποθήκευση γραφικών για εξωτερική χρήση

Οι συναρτήσεις γραφικών παραστάσεων, όπως η plot, μας δίνουν το αποτέλεσμά τους σε ένα εσωτερικό παράθυρο στο R. Για να χρησιμοποιήσουμε το γραφικό αποτέλεσμα εξωτερικά, π.χ. να το ενσωματώσουμε σε μια έκθεση σε αρχείο κειμένου (Word .docx), πρέπει προηγουμένως να το αποθηκεύσουμε σε κάποιο αρχείο γραφικών. Το R δίνει πολλές τέτοιες δυνατότητες. Για παράδειγμα, μπορούμε να δημιουργήσουμε ένα αρχείο γραφικών τύπου jpg ως εξής:

```
> jpeg("test.jpg")
> plot(d)
> dev.off()
```

Για αρχείο τύπου pdf χρησιμοποιούμε, αντίστοιχα, τις εξής συναρτήσεις:

```
> pdf("test.pdf")
> plot(d)
> dev.off()
```

Γενικά, η διαδικασία περιλαμβάνει τρία βήματα. Στο πρώτο βήμα ανοίγουμε ένα εξωτερικό αρχείο για να υποδεχτεί το γραφικό μας. Ανάλογα με τη συνάρτηση ρυθμίζεται το είδος του αρχείου: η συνάρτηση jpeg ανοίγει αρχείο τύπου jpg, η συνάρτηση pdf ανοίγει αρχείο τύπου pdf κ.ο.κ. Το όνομα του αρχείου είναι στη διακριτική μας ευχέρεια και μπορούμε να το βαφτίσουμε όπως μας αρέσει—στο παραπάνω παράδειγμα ονομάζεται «test». Η κατάληξη του ονόματος πρέπει να συμφωνεί με τον τύπο του αρχείου, δηλαδή, για τη συνάρτηση jpeg πρέπει να είναι «.jpg», ενώ για τη συνάρτηση pdf πρέπει να είναι «.pdf».

Στο δεύτερο βήμα εκτελούμε τη διαδικασία παραγωγής του γραφικού, είτε με τη συνάρτηση plot είτε με οποιαδήποτε άλλη (π.χ. barplot, boxplot, hist, histnorm κλπ.). Οτιδήποτε κάνουμε σε αυτό το βήμα δεν θα εμφανιστεί στο εσωτερικό παράθυρο του R αλλά θα κατευθυνθεί στο εξωτερικό αρχείο που έχει ανοίξει. Οπότε για να είμαστε σίγουροι για το αποτέλεσμα θα πρέπει προηγουμένως να το έχουμε δοκιμάσει εσωτερικά.

Στο τρίτο βήμα ολοκληρώνουμε τη δημιουργία του εξωτερικού αρχείου με τη συνάρτηση dev.off(). Η συνάρτηση αυτή είναι η ίδια ανεξαρτήτως από τον τύπο του αρχείου που έχουμε ανοίξει. Το βήμα αυτό είναι απολύτως απαραίτητο—αν το παραλείψουμε τότε το εξωτερικό αρχείο δεν θα μπορεί να ανοιχτεί από κανένα πρόγραμμα.

Με αυτή τη διαδικασία τριών βημάτων αποθηκεύεται ένα αρχείο γραφικών στον επιλεγμένο φάκελο εργασίας στον υπολογιστή μας. Από εκεί μπορούμε να το ανοίξουμε με οποιοδήποτε πρόγραμμα αναγνωρίζει γραφικά, ή απλώς να το τραβήξουμε πάνω στο κείμενό μας στο word, οπότε και θα εμφανιστεί η γραφική παράσταση μέσα στο κείμενο, ως εικόνα.

Ε. Προσδιορισμός σχέσεων μεταβλητών στις συναρτήσεις

Πολλές συναρτήσεις, στις οποίες έχει νόημα η διάκριση μεταξύ ανεξάρτητης και εξαρτημένης μεταβλητής, δέχονται ορίσματα με μορφή «μαθηματικού τύπου» (formula). Η μορφή αυτή χαρακτηρίζεται από το σύμβολο ~ (συνήθως βρίσκεται στο πλήκτρο αριστερά από τον αριθμό 1 στο πληκτρολόγιο). Αριστερά από το σύμβολο αυτό τοποθετείται η εξαρτημένη μεταβλητή και δεξιά η ανεξάρτητη (ή ανεξάρτητες, αν είναι πάνω από μία).

Για παράδειγμα, στη συνάρτηση t.test, αντί για δύο ξεχωριστά διανύσματα (όπως στο παραπάνω παράδειγμα), μπορούμε να δώσουμε απευθείας τη σχέση που μας ενδιαφέρει από το πλαίσιο δεδομένων. Χρησιμοποιώντας τις μεταβλητές x (ποσοτική/numeric) και f (ποιοτική/factor) όπως ορίστηκαν παραπάνω, στο πλαίσιο δεδομένων d, θα γράφαμε:

```
> t.test(x~f,d)
```

Το ίδιο μπορούμε να κάνουμε και στη συνάρτηση plot:

```
> plot(x~f,d)
```

Με τον τρόπο αυτό προσδιορίζουμε εμείς ποια θέλουμε να είναι η μεταβλητή στον οριζόντιο άξονα (η ανεξάρτητη) και ποια στον κατακόρυφο (εξαρτημένη), όποια κι αν είναι η θέση τους μέσα στο πλαίσιο δεδομένων. Εννοείται ότι η μορφή αυτή μπορεί να χρησιμοποιηθεί και όταν υπάρχουν κι άλλες μεταβλητές μέσα στο ίδιο πλαίσιο δεδομένων, οι οποίες έτσι θα αγνοηθούν.

Στ. Βοήθεια

Το R διαθέτει πάρα πολλές συναρτήσεις σε πολλές διαφορετικές βιβλιοθήκες. Είναι αδύνατο να θυμάται κανείς απέξω όλες τις προδιαγραφές για τη χρήση κάθε συνάρτησης και τις λεπτομέρειες των αποτελεσμάτων. Για το σκοπό αυτό υπάρχει δυνατότητα βοήθειας, πληκτρολογώντας ένα (λατινικό) ερωτηματικό πριν από το όνομα της συνάρτησης. Για παράδειγμα, για να δούμε πληροφορίες σχετικά με τη χρήση της συνάρτησης mean, δίνουμε

```
> ?mean
```

Αν δεν γνωρίζουμε ακριβώς το όνομα της συνάρτησης που χρειαζόμαστε, μπορούμε να αναζητήσουμε συναρτήσεις σχετικά με την ανάλυση που μας ενδιαφέρει, εξετάζοντας όλες τις βιβλιοθήκες που έχουμε εγκαταστήσει στον υπολογιστή μας. Αυτό γίνεται πληκτρολογώντας δύο κολλητά ερωτηματικά αντί για ένα. Για παράδειγμα, αν χρειαζόμαστε πληροφορίες σχετικά με αναλύσεις συσχέτισης (correlation), δίνουμε

```
> ??correlation
```

Το R εξετάζει όλες τις εγκατεστημένες βιβλιοθήκες και μας επιστρέφει κατάλογο συναρτήσεων.

Αθανάσιος Πρωτόπαπας
Ιούνιος 2012