

# Το T<sub>E</sub>X στο χώρο της στατιστικής: η δύναμη του ελεύθερου λογισμικού

---

Ιωάννης Κ. Δημάκος

*Πανεπιστήμιο Πατρών*

*Παιδαγωγικό Τμήμα Δημοτικής Εκπαίδευσης*

*Τομέας Ψυχολογίας*

*265 00 Πάτρα*

*URL: [www.elemedu.upatras.gr/dimakos](http://www.elemedu.upatras.gr/dimakos)*

*H/T: [idimakos@upatras.gr](mailto:idimakos@upatras.gr)*

Στο παρόν άρθρο, παρουσιάζεται σύντομα το πρόγραμμα και το περιβάλλον στατιστικού προγραμματισμού R. Αυτό το ανοικτό περιβάλλον, σε συνδυασμό με την ευκολία, την ευελιξία και τις ικανότητες του T<sub>E</sub>X και των συναφών προγραμμάτων του, παρέχει την δυνατότητα στο συντάκτη να δημιουργήσει κείμενα με μαθηματικές και στατιστικές αναλύσεις (και τα απαραίτητα συνοδευτικά γραφικά) υψηλής πιστότητας.

**T<sub>E</sub>X in the field of Statistics: The power of free software**, by Ioannis K. Dimakos — The statistical programming environment known as R, a free statistical software, is presented in this article. This environment, in conjunction with the flexibility, ease of use and capabilities of T<sub>E</sub>X and its associated programs, offers every author the ability to create high fidelity mathematical and statistical texts (along with the necessary graphics).

## 1 Εισαγωγή

Θα ήταν κοινοτοπία να αναφερθούμε για άλλη μια φορά στη δύναμη και τις διάφορες χρήσεις του ελεύθερου λογισμικού. Στον επιστημονικό χώρο υπάρχουν πολλά τέτοια παραδείγματα ελεύθερου κώδικα που συνυπάρχουν μαζί με αντίστοιχα εμπορικά πακέτα με ίδιες ή παραπλήσιες δυνατότητες. Μια έρευνα στο γνωστό διαδικτυακό τόπο [www.google.com](http://www.google.com) χρησιμοποιώντας τους όρους *free scientific software*, αναδεικνύει μερικές δεκάδες εκατομμύρια αποτελέσματα. Μάλιστα, τα περισσότερα από τα προγράμματα αυτά διαθέτουν εκδόσεις για διάφορα λειτουργικά συστήματα (Unix και τα παράγωγά του, MS-Windows, Mac OS). Στο άρθρο αυτό θα παρουσιάσουμε ένα σχετικά καινούριο πρόγραμμα στατιστικής ανάλυσης και προγραμματισμού, το R [1] το οποίο διανέμεται μέσω της γνωστής άδειας ελεύθερου λογισμικού GPL και θα εστιάσουμε στις δυνατότητες συνεργασίας του προγράμματος αυτού με ολόκληρη την οικογένεια του T<sub>E</sub>X.

## 2 Το παράδειγμα της στατιστικής

Στο χώρο της στατιστικής, ένα από τα παλαιότερα προγράμματα στατιστικής ανάλυσης είναι και το γνωστό σε σχεδόν όλους μας SPSS<sup>1</sup>. Το πρόγραμμα αυτό είναι ένα πλήρες «πακέτο» εισαγωγής, επεξεργασίας, ανάλυσης και παρουσίασης πληροφοριών και δεδομένων<sup>2</sup>. Όπως συνέβη και με άλλα ομοειδή προγράμματα, το SPSS αν και ξεκίνησε από τον ακαδημαϊκό χώρο ως πρόγραμμα στατιστικής επεξεργασίας, στη συνέχεια μετεξελιχθηκε σε ένα από τα κορυφαία προγράμματα ανάλυσης παντός είδος πληροφοριών.

Αν και είναι βέβαιο πως το πρόγραμμα αυτό, όπως και άλλα προγράμματα της αυτής κατηγορίας, ικανοποιεί πλήρως όσους το χρησιμοποιούν, ωστόσο έχει ορισμένα μειονεκτήματα για ένα σημαντικό αριθμό χρηστών. Δεν αναφέρομαι, βέβαια, σε θέματα αλγοριθμικής ακρίβειας των προγραμμάτων αυτών. Στο διαδικτυακό τόπο [www.nist.gov](http://www.nist.gov), αλλά και αλλού, υπάρχουν αναφορές και βάσεις δεδομένων για τη διαπίστωση της στατιστικής και αλγοριθμικής ακρίβειας διαφόρων προγραμμάτων. Για περισσότερες πληροφορίες, οι McCullough και Wilson [2] μελέτησαν την ακρίβεια των αλγορίθμων και σύγκριναν την επάρκεια διαφόρων στατιστικών προγραμμάτων (μεταξύ αυτών και του module στατιστικής ανάλυσης του γνωστού MS-Excel, για το οποίο η κρίση δεν ήταν καθόλου ικανοποιητική και το οποίο δεν ενδείκνυται για στατιστικές αναλύσεις).

Τα μειονεκτήματα στα οποία θα αναφερθώ εγώ είναι άλλα. Πρώτον, τα περισσότερα εμπορικά προγράμματα-πακέτα είναι ακριβά και απαιτούν ετήσιες άδειες ανανέωσης που κοστίζουν αρκετά μεγάλα ποσά. Επίσης, τα προγράμματα αυτά είναι «κλειστά», με άλλα λόγια, δεν γνωρίζει ο χρήστης, και αρκετές φορές δεν είναι εύκολο να βρει, τους αλγόριθμους που χρησιμοποιούνται στις διάφορες στατιστικές αναλύσεις. Παράλληλα, στα πλαίσια της διευκόλυνσης ενός αρκετά μεγάλου και ετερογενούς ακροατηρίου, τα μενού των επιλογών στατιστικής ανάλυσης περιορίζουν εν μέρει κάποιους χρήστες που θα ήθελαν να παρέμβουν και να επιχειρήσουν μια πιο σύνθετη ανάλυση. Αυτό το τελευταίο χαρακτηριστικό τα καθιστά συχνά ανεπαρκή για τη διδασκαλία της Στατιστικής και της Μεθοδολίας της Έρευνας. Μετατρέπεται έτσι η διδασκαλία των σχετικών μαθημάτων σε διδασκαλία των επιλογών ενός προγράμματος στατιστικής ανάλυσης. Θα προσπαθήσω, στη συνέχεια, να αναπτύξω και να απαντήσω στα μειονεκτήματα αυτά παρουσιάζοντας το πρόγραμμα R.

## 3 Το πρόγραμμα R

Σύμφωνα με τον επίσημο δικτυακό τόπο του προγράμματος, [www.r-project.org](http://www.r-project.org), το R βασίζεται στη γλώσσα προγραμματισμού S που δημιουργήθηκε από τον John Chambers και τους συνεργάτες του στα εργαστήρια Bell της AT&T. Το R

<sup>1</sup>Εξίσου παλαιά προγράμματα στατιστικής πρέπει να θεωρηθούν τα προγράμματα BMDP και SAS.

<sup>2</sup>Δεν αναφέρομαι σε μεμονωμένους αλγόριθμους και προγράμματα τα οποία έχουν γραφτεί για συγκεκριμένες στατιστικές αναλύσεις, αλλά για ολοκληρωμένα προγράμματα. Τέτοιοι μεμονωμένοι αλγόριθμοι και μικρά προγράμματα υπάρχουν ακόμα στο δικτυακό χώρο του πανεπιστημίου Carnegie Mellon στη διεύθυνση: [www.statlib.cmu.edu](http://www.statlib.cmu.edu).

θεωρείται και *γλώσσα* και *περιβάλλον στατιστικού προγραμματισμού* και είναι προϊόν των Robert Gentleman και Ross Ihaka[3].

Το R ακολουθεί έναν εξαμηνιαίο κύκλο διανομής. Νέες εκδόσεις του προγράμματος κυκλοφορούν κάθε Οκτώβριο και Απρίλιο περίπου. Οι εκδόσεις είναι αριθμημένες (της μορφής  $x.y.z$ ), με τους  $x.y$  να αφορούν την κύρια έκδοση (major release), και το  $z$  τις όποιες μικρές διορθώσεις (minor release). Παράδειγμα, η έκδοση του Οκτωβρίου 2007 είχε κωδικό αριθμό 2.6.2, ενώ πριν λίγο κυκλοφόρησε η έκδοση 2.7.0. Πάντοτε κυκλοφορούν εκδόσεις πηγαίες (source code) και εκτελέσιμες (binary release) για διάφορα λειτουργικά συστήματα.

Τον κώδικα του R επιμελείται για λογαριασμό του R Foundation μια κεντρική ομάδα προγραμματιστών, η οποία δέχεται μέσω διαφόρων forum ηλεκτρονικής επικοινωνίας με την υπόλοιπη κοινότητα χρηστών του προγράμματος προτάσεις για βελτιώσεις, προσθήκες, επισημάνσεις για λάθη, παραλείψεις, κ.λπ. Κάθε εγκατάσταση του R προσφέρει βασικές στατιστικές και αλγοριθμικές δυνατότητες στο χρήστη, αλλά και μια δυνατή μηχανή παραγωγής γραφημάτων τα οποία ο χρήστης μπορεί να αποθηκεύσει ως αρχεία διαφορετικών τύπων, όπως: Windows metafile, bitmap, PostScript, jpeg, png, pdf, κ.λπ. Μάλιστα, στο περιβάλλον των MS-Windows ο χρήστης με το γνωστό «δεξί κλικ» μπορεί να αποθηκεύσει τα παραγόμενα γραφικά, ενώ σε άλλα λειτουργικά συστήματα, ο χρήστης πρέπει να ορίσει τα χαρακτηριστικά του παραγόμενου αρχείου γραφικών.

Το R έρχεται με μια σειρά πακέτων (packages) που προσφέρουν βασικές στατιστικές λειτουργίες. Τα πακέτα αυτά τα καλεί ο χρήστης από τη γραμμή εντολών του προγράμματος (το R δεν διαθέτει κάποιο ξεχωριστό γραφικό περιβάλλον (GUI) και μενού εντολών για τις στατιστικές αναλύσεις) και στη συνέχεια έχει πρόσβαση στις ρουτίνες του κάθε πακέτου<sup>3</sup>. Εκτός όμως από τα βασικά πακέτα που διανέμονται με το πρόγραμμα, η κοινότητα του R έχει συνεισφέρει συνολικά 1363 πακέτα που επεκτείνουν τις ήδη ευρύτατες δυνατότητες του προγράμματος με εφαρμογές στις κοινωνικές επιστήμες, στην οικονομική ανάλυση, στη φασματοσκοπία, στη ανάλυση γονιδιωμάτων και αλλού. Οι εφαρμογές είναι πραγματικά απεριόριστες.

Όλα τα προαναφερθέντα προγράμματα και πακέτα υπάρχουν διαθέσιμα σε ένα παγκόσμιο δίκτυο τοπικών εξυπηρετητών. Όπως και το  $\TeX$ , έτσι και το R διαθέτει το δικό του Comprehensive R Archive Network (CRAN), με αρκετούς τοπικούς κόμβους (mirrors) ανά τον κόσμο. Η διεύθυνσή του είναι η εξής: [cran.r-project.org](http://cran.r-project.org). Στο CRAN ο χρήστης του R μπορεί να βρει τα επίσημα εγχειρίδια της κάθε έκδοσης (που συνοδεύουν ούτως ή άλλως κάθε πακέτο εγκατάστασης, πηγαίο ή εκτελέσιμο), εγχειρίδια που έχουν γραφτεί από χρήστες του R σε διαφορετικές γλώσσες (εκτός της Αγγλικής) και μπορεί να αφορούν μια συγκεκριμένη χρήση του προγράμματος. Μέσα από το πρόγραμμα, ο χρήστης μπορεί να συνδεθεί με κάποιο τοπικό κόμβο του CRAN και να εμπλουτίσει τη βιβλιοθήκη των πακέτων με νέα πακέτα ή να ζητήσει τον έλεγχο για την ανανέωση των ήδη υπάρχοντων πακέτων.

<sup>3</sup>Εκ κατασκευής το R δεν έχει μενού εντολών, έχει όμως πακέτα, τα οποία του δίνουν αυτή τη δυνατότητα. Για παράδειγμα, δύο πακέτα είναι το JGR και το Rcmdr τα οποία καλύπτουν επάξια το κενό αυτό.

## 4 Το R και η οικογένεια T<sub>E</sub>X

Όμως, σκοπός του παρόντος άρθρου είναι να τονίσει τη δυνατότητα συνεργασίας δύο ελεύθερων προγραμμάτων: του R και του T<sub>E</sub>X. Για να γίνει όμως κατανοητή η συνεργασία αυτή, πρέπει πρώτα να αναφερθούμε πρώτα στον τρόπο λειτουργίας του R.

Όπως προαναφέρθηκε, το R δεν έχει κάποιο επιτηδευμένο γραφικό περιβάλλον. Μάλιστα όταν καλεί ο χρήστης το πρόγραμμα (είτε σε περιβάλλον MS-Windows είτε σε περιβάλλον Unix ή παραγώγων του), το πρόγραμμα εκτελείται και παρουσιάζει στο χρήστη τα επόμενα πληροφοριακά μηνύματα σχετικά με την έκδοση, την ημερομηνία δημιουργίας της συγκεκριμένης έκδοσης, και επίσης ορισμένες πληροφορίες σχετικά με το πώς μπορεί να ζητήσει βοήθεια ή να δει μερικές από τις δυνατότητες του προγράμματος. Στη συνέχεια, το πρόγραμμα περιμένει τις εντολές του χρήστη. Κάθε εντολή συνοδεύεται υποχρεωτικά από παρενθέσεις (αν δεν μπου οι παρενθέσεις, τότε το πρόγραμμα εμφανίζει τον κώδικα της εντολής, στοιχείο που αναδεικνύει και τον «ανοικτό» χαρακτήρα του προγράμματος).

R version 2.6.2 (2008-02-08)

Copyright (C) 2008 The R Foundation for Statistical Computing  
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.

>

Στο prompt, ο χρήστης μπορεί να κάνει μια απλή ή σύνθετη αριθμητική πράξη, να καλέσει δεδομένα, πακέτα, να ζητήσει στατιστικές αναλύσεις. Για παράδειγμα: πληκτρολογώντας

```
1 + 2
```

το πρόγραμμα απαντάει:

```
[1] 3
```

Ομοίως, ο χρήστης μπορεί να ζητήσει:

```
a <- 1
```

Το *TEX* στο χώρο της στατιστικής

41

```
b <- 2
c <- a + b
```

Το πρόγραμμα απαντάει ως εξής:

```
>
```

Γιατί; Απλά το πρόγραμμα ανέθεσε την τιμή 1 στο αντικείμενο (object) *a*, την τιμή 2 στο αντικείμενο *b* και την τιμή που προκύπτει από το άθροισμα των τιμών των δύο αυτών αντικειμένων στο νέο αντικείμενο *c*. Αν θέλει ο χρήστης να δει το αποτέλεσμα της πράξης, δηλαδή το αντικείμενο *c*, πρέπει να το καλέσει:

```
> c
[1] 3
```

Το πρόγραμμα μας δίνει τη δυνατότητα να εξερευνήσουμε περισσότερο το αντικείμενο *c*.

```
> str(c)
num 3
```

Με άλλα λόγια, μαθαίνουμε ότι το αντικείμενο είναι αριθμητικής φύσεως (numeric) και έχει την τιμή 3.

Η δυνατότητα να δουλεύουμε με αντικείμενα είναι πολύ χρήσιμη όταν το αντικείμενο είναι μία λίστα (list) ή ένα σετ δεδομένων (data frame). Ας υποθέσουμε ότι χρησιμοποιούμε το σετ δεδομένων *warpbreaks* που συμπεριλαμβάνεται στο πακέτο *datasets* του προγράμματος. Πρώτα φορτώνουμε το σετ δεδομένων στο χώρο εργασίας του προγράμματος με την εντολή *data(warpbreaks)*. Στη συνέχεια μπορούμε να εξετάσουμε το περιεχόμενο του αντικειμένου *warpbreaks* με την εντολή *str(warpbreaks)*.

```
> data(warpbreaks)
> str(warpbreaks)
'data.frame': 54 obs. of 3 variables:
 $ breaks : num 26 30 54 25 70 52 51 26 67 18 ...
 $ wool : Factor w/ 2 levels "A","B": 1 1 1 1 1 1 1 1 1 1 ...
 $ tension: Factor w/ 3 levels "L","M","H": 1 1 1 1 1 1 1 1 1 2 ...
> summary(warpbreaks)
      breaks      wool      tension
Min.   :10.00   A:27   L:18
1st Qu.:18.25   B:27   M:18
Median :26.00           H:18
Mean   :28.15
3rd Qu.:34.00
Max.   :70.00
>
```

Πληροφορούμαστε ότι το σετ δεδομένων *warpbreaks* έχει 54 παρατηρήσεις και 3 μεταβλητές, τις εξής:

**breaks** : ποσοτική μεταβλητή,

**wool** : κατηγορική μεταβλητή με 2 επίπεδα,

**tension** : επίσης κατηγορική μεταβλητή με 3 επίπεδα.

Ακόμα μία χρήσιμη εντολή είναι και η `summary(warpbreaks)` που μας δίνει την περίληψη των 5 στιγμών (ελάχιστη τιμή, 25ο εκατοστημόριο, διάμεσος, 75ο εκατοστημόριο, μέγιστη τιμή) και το μέσο όρο για την ποσοτική μεταβλητή, ενώ για τις κατηγορικές μεταβλητές μας δίνει τις κατανομές των τιμών ανά κατηγορία.

Αναλύοντας τα δεδομένα αυτά με τη μέθοδο της ANOVA: Analysis of Variance (ανάλυσης της διακύμανσης, με δύο παράγοντες μεταξύ ομάδων) παίρνουμε τα ακόλουθα:

```
> aov(breaks~wool*tension,data=warpbreaks)
```

Call:

```
aov(formula = breaks ~ wool * tension, data = warpbreaks)
```

Terms:

	wool	tension	wool:tension	Residuals
Sum of Squares	450.667	2034.259	1002.778	5745.111
Deg. of Freedom	1	2	2	48

Residual standard error: 10.94028

Estimated effects may be unbalanced

Τα αποτελέσματα αυτά δεν είναι πλήρη. Πληροφορούμαστε ότι έγινε ανάλυση της διακύμανσης της μεταβλητής `breaks` με δύο παράγοντες (`wool` και `tension`), από το σετ δεδομένων `warpbreaks`. Στη συνέχεια, το πρόγραμμα παρουσιάζει μερικά βασικά στοιχεία της ανάλυσης. Για να εμφανιστούν τα αποτελέσματα με μια καλύτερη μορφή, πρέπει να δοθεί μια άλλη εντολή, η `summary()`:

```
> summary(aov(breaks~wool*tension,data=warpbreaks))
```

Βλέπουμε ότι μπορούμε να ενσωματώσουμε μια εντολή μέσα σε μια άλλη. Στην περίπτωση αυτή, η δεύτερη εντολή αποτελεί το όρισμα της πρώτης. Εναλλακτικά, μπορούμε να επιλέξουμε να αποθηκεύσουμε τα αποτελέσματα σε αντικείμενο και να δώσουμε τη νέα εντολή στο αντικείμενο αυτό:

```
> breaks.results <- aov(breaks~wool*tension,data=warpbreaks)
```

```
> summary(breaks.results)
```

Όπως και να έχει, το πρόγραμμα θα μας δώσει ένα κατατοπιστικότερο πίνακα αποτελεσμάτων:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
wool	1	450.7	450.7	3.7653	0.0582130 .
tension	2	2034.3	1017.1	8.4980	0.0006926 ***
wool:tension	2	1002.8	501.4	4.1891	0.0210442 *

Το T<sub>E</sub>X στο χώρο της στατιστικής

43

```
Residuals    48 5745.1  119.7
```

```
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Τα αποτελέσματα με την παρούσα μορφή τους είναι περισσότερο διαφωτιστικά, αφού το πρόγραμμα μας δίνει το γνωστό πίνακα της ανάλυσης της διακύμανσης με τους σχετικούς ανεξάρτητους παράγοντες, τους βαθμούς ελευθερίας (df) που αντιστοιχούν σε κάθε παράγοντα, τα αθροίσματα τετραγώνων των παραγόντων (Sums of Squares), τα μέσα αθροίσματα τετραγώνων (Mean Squares), την τιμή του κριτηρίου F της ανάλυσης της διακύμανσης και την πιθανότητα που συνοδεύει κάθε αποτέλεσμα. Στη συνέχεια ο ερευνητής μπορεί να καλέσει άλλες εντολές για τη δημιουργία γραφικών αναπαραστάσεων (βλ. Εικόνα 1) ή να πραγματοποιήσει επιμέρους συγκρίσεις με το κριτήριο του Tukey.

```
> interaction.plot(tension,wool,breaks)
> TukeyHSD(breaks.results,"tension")
> plot(TukeyHSD(breaks.results,"tension"))
```

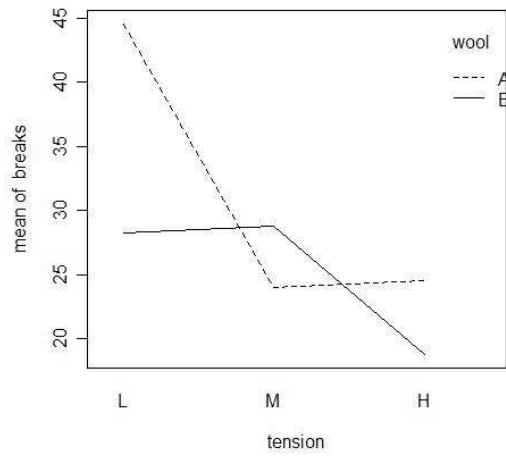
Η πρώτη εντολή θα δημιουργήσει το γράφημα της Εικόνας 1, ενώ η τρίτη εντολή το γράφημα της Εικόνας 2. Η δεύτερη εντολή θα δώσει τα παρακάτω αποτελέσματα των επιμέρους συγκρίσεων για έναν από τους δύο ανεξάρτητους παράγοντες της ανάλυσης:

```
Tukey multiple comparisons of means
 95% family-wise confidence level
```

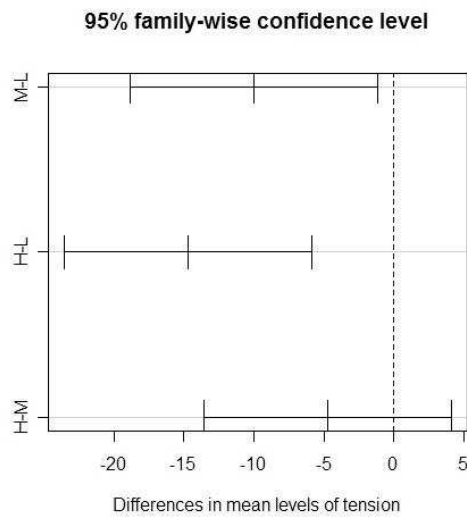
```
Fit: aov(formula = breaks ~ tension * wool, data = warpbreaks)
```

```
$tension
      diff      lwr      upr      p adj
M-L -10.000000 -18.81965 -1.180353 0.0228554
H-L -14.722222 -23.54187 -5.902575 0.0005595
H-M  -4.722222 -13.54187  4.097425 0.4049442
```

Όμως, είπαμε ότι το R μπορεί να συνεργαστεί με την οικογένεια προγραμμάτων του T<sub>E</sub>X και ακόμα δεν έχουμε δείξει πώς γίνεται αυτό. Από τη μικρή παρουσίαση των στατιστικών δυνατοτήτων του R φάνηκε ότι το πρόγραμμα μπορεί να παρουσιάσει τα αποτελέσματα των αναλύσεων είτε με τη μορφή κειμένου είτε γραφικών. Πώς όμως τα αποτελέσματα αυτά θα εισαχθούν στο σώμα μιας τεχνικής αναφοράς, ενός επιστημονικού άρθρου; Εδώ μπαίνει το T<sub>E</sub>X και αρκετά από τα πακέτα που έχουν συνεισφέρει οι χρήστες του προγράμματος R. Πιο συγκεκριμένα, στο CRAN υπάρχουν διαθέσιμα πακέτα τα οποία μπορούν να μετατρέψουν τα αποτελέσματα του προγράμματος σε μορφή που να διαβάζεται από το T<sub>E</sub>X. Ένα τέτοιο πακέτο είναι και το πακέτο xtable, το οποίο μπορεί να πάρει τον πίνακα της ανάλυσης της διακύμανσης και να τον κωδικοποιήσει σε μορφή τέτοια ώστε να μπορεί να εισαχθεί στο κείμενο με τις γνωστές εντολές για τη δημιουργία πινάκων. Έτσι, δίνοντας τις εντολές:



Εικόνα 1: Απεικόνιση της ανάλυσης.



Εικόνα 2: Απεικόνιση των διαφορών μεταξύ των επιπέδων του ενός παράγοντα.

Το  $\TeX$  στο χώρο της στατιστικής

45

```
> library(xtable)
> xtable(breaks.results)
```

το πρόγραμμα θα εμφανίσει το επόμενο κομμάτι κώδικα:

```
% latex table generated in R 2.6.2 by xtable 1.5-2 package
% Fri Apr 18 02:34:32 2008
\begin{table}[ht]
\begin{center}
\begin{tabular}{lrrrrr}
\hline
& Df & Sum Sq & Mean Sq & F value & Pr(>F) \\
\hline
wool & 1 & 450.67 & 450.67 & 3.77 & 0.0582 \\
tension & 2 & 2034.26 & 1017.13 & 8.50 & 0.0007 \\
wool:tension & 2 & 1002.78 & 501.39 & 4.19 & 0.0210 \\
Residuals & 48 & 5745.11 & 119.69 & & \\
\hline
\end{tabular}
\end{center}
\end{table}
```

Φυσικά, ο κώδικας αυτός είναι έτοιμος για εισαγωγή στο κείμενο που επεξεργαζόμαστε και θα μας δώσει τον πίνακα που ακολουθεί:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
wool	1	450.67	450.67	3.77	0.0582
tension	2	2034.26	1017.13	8.50	0.0007
wool:tension	2	1002.78	501.39	4.19	0.0210
Residuals	48	5745.11	119.69		

Όμως, δεν είναι μόνο αυτό. Ο συνδυασμός δύο ισχυρών προγραμμάτων μπορεί να αποδώσει περισσότερα, ειδικά όταν εισερχόμαστε στο χώρο του *literate programming*, ή λογοτεχνικού προγραμματισμού. Το R μπορεί να επεξεργαστεί αρχεία που περιέχουν εντολές για το πρόγραμμα και στη συνέχεια να δημιουργήσει αρχεία  $\LaTeX$  που περιέχουν τα αποτελέσματα των εντολών αυτών.

Πιο συγκεκριμένα, δημιουργούμε το ακόλουθο αρχείο με κατάληξη *.rnw*. Στο αρχείο αυτό θα βάλουμε μεικτό κώδικα  $\LaTeX$  και R.

```
\documentclass[a4paper]{article}
\begin{document}
<<>>=
data(sleep)
t.test(extra~group,data=sleep,paired=TRUE)
```

```
@
\begin{center}
<<fig =TRUE , echo =FALSE >>=
plot(extra~group,data=sleep)
@
\end{center}
\end{document}
```

Για τις ανάγκες του παραδείγματος χρησιμοποιήσαμε ένα κλασικό σετ δεδομένων από το 1905 του William T. Student, δημιουργού του t-test. Μετά τον αρχικό κώδικα L<sup>A</sup>T<sub>E</sub>X, ακολουθούν οι εντολές για να καλέσουμε τα δεδομένα και να ζητήσουμε το κριτήριο t-test για τη σύγκριση δυο εξισωμένων δειγμάτων. Στη συνέχεια και εντός πλαισίου που ορίζεται από το σύμβολο @, ζητάμε τη δημιουργία σχήματος (χωρίς την αναπαραγωγή του σχετικού κώδικα). Επίσης, ζητάμε το σχήμα αυτό να είναι κεντραρισμένο. Όταν εκτελέσουμε τον κώδικα αυτό με τη βοήθεια του προγράμματος R, θα πάρουμε ένα αρχείο με το ίδιο όνομα όπως το πηγαίο αλλά με κατάληξη .tex και τα αρχεία των γραφικών (σε μορφή encapsulated PostScript και pdf. Το αρχείο με τον κώδικα για το L<sup>A</sup>T<sub>E</sub>X είναι το ακόλουθο:

```
\documentclass[a4paper]{article}
\usepackage{C:/PROGRA~1/r/share/texmf/Sweave}
\begin{document}
\begin{Schunk}
\begin{Sinput}
> data(sleep)
> t.test(extra ~ group, data = sleep, paired = TRUE)
\end{Sinput}
\begin{Soutput}
Paired t-test

data:  extra by group
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.4598858 -0.7001142
sample estimates:
mean of the differences
                -1.58
\end{Soutput}
\end{Schunk}
\begin{center}
\includegraphics{eutypon-r-sweave2-002}
\end{center}

\end{document}
```

Ο κώδικας του προγράμματος R εισάγεται σε ειδικό περιβάλλον με το όνομα Schunk. Οι εντολές του προγράμματος και τα αποτελέσματα των εντολών εισάγονται μέσα στο περιβάλλον Sinput και Soutput, αντίστοιχα. Στη συνέχεια, με τη σχετική εντολή παράγεται το γράφημα που έχουμε ζητήσει.

Η συνεργασία του R με την ευρύτερη οικογένεια του T<sub>E</sub>X δεν περιορίζεται μόνο στα προαναφερθέντα. Ένα πρόγραμμα σαν και το R δεν θα μπορούσε παρά να έχει τη σχετική βιβλιογραφία με μορφή .texi. Η μορφή αυτή μπορεί κατόπιν επεξεργασίας να γίνει dvi, αλλά και pdf ή info. Ακόμα και η βιβλιογραφία που συνοδεύει κάθε πακέτο εντολών είναι σε μορφή L<sup>A</sup>T<sub>E</sub>X.

## 5 Αντί επιλόγου

Το παρόν άρθρο δεν μπορεί να είναι, ούτε και φιλοδοξούσε να είναι, ένα εισαγωγικό tutorial της γλώσσας και του περιβάλλοντος στατιστικού προγραμματισμού R. Ο ενδιαφερόμενος χρήστης μπορεί, και πρέπει, να ανατρέξει στην επίσημη ιστοσελίδα του προγράμματος στη διεύθυνση [www.r-project.org](http://www.r-project.org) ή στον πλησιέστερο κόμβο του CRAN για την πληρέστερη δυνατή πληροφόρηση και ενημέρωσή του.

Δυο ήταν οι στόχοι του άρθρου αυτού. Πρώτον, να ενημερώσει για ένα νέο και συναρπαστικό πρόγραμμα στατιστικού προγραμματισμού. Φυσικά κάτι τέτοιο δεν είναι δυνατό μέσα από τις λίγες σελίδες του παρόντος άρθρου. Όμως, μέσω των παραδειγμάτων που παρουσιάστηκαν ο αναγνώστης ίσως να μπορέσει να εξάγει ορισμένα χρήσιμα συμπεράσματα για τις δυνατότητες του R.

Δεύτερον, να παρουσιάσει τις δυνατότητες του προγράμματος αυτού σε συνδυασμό με την οικογένεια προγραμμάτων του T<sub>E</sub>X. Το R χρειάζεται το T<sub>E</sub>X για τη βιβλιογραφία του, την τεκμηρίωση των εντολών και των δεδομένων και των παραδειγμάτων που εμπεριέχονται σε κάθε διανομή. Επιπλέον, στα πλαίσια του λογοτεχνικού προγραμματισμού, ο χρήστης του προγράμματος R μπορεί να ενσωματώσει εντολές από τα δύο αυτά προγράμματα και να γράψει αναλυτικά κείμενα που επεξηγούν κάθε βήμα του κώδικα. Οι δυνατότητες είναι πραγματικά απεριόριστες.

## Αναφορές

- [1] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008. (ULR: <http://www.R-project.org>.)
- [2] B. McCullough and B. Wilson, “On the accuracy of statistical procedures in Microsoft Excel 97.” *Computational Statistics & Data Analysis*, vol. 31 (1999), no 1, pp. 27–37.
- [3] R. Ihaka and R. Gentleman, “R: A language for data analysis and graphics.” *Journal of Computational and Graphical Statistics*, vol. 5 (1996), no 3, pp. 299–314.