



Ανάλυση Δεδομένων με Η/Υ- 5 Μαρτίου 2012

Άσκηση 1 (30 μονάδες)

```

PRun<-function(k,N){
  x<-c(rep(1,k-1),1-1/2^(k-1))
  i<-k
  while (i<N) {x<-c(x[-1],x[k]-x[1]/2^k)
             i<-i+1}
  return(x[1:min(k,N)])
}

```

Στρίβουμε N φορές ένα νόμισμα και ορίζουμε ως $\ell(N)$ το μήκος της μεγαλύτερης ακολουθίας από διαδοχικά ίδια αποτελέσματα. Π.χ. αν $N = 6$, στο ενδεχόμενο $\{K, \Gamma, K, K, K, \Gamma\}$ έχουμε $\ell(6) = 3$, ενώ στο ενδεχόμενο $\{\Gamma, \Gamma, K, \Gamma, K, K\}$ έχουμε $\ell(6) = 2$. Αν $k, N \in \mathbb{N}$, η διπλανή συνάρτηση PRun επιστρέφει την πιθανότητα $\mathbb{P}[\ell(N) < k]$.

α) Τρέξτε με το χέρι την PRun για να υπολογίσετε την πιθανότητα $\mathbb{P}[\ell(7) < 4]$. Δείξτε όλες τις ενδιαμέσες τιμές που παίρνει το διάνυσμα x .

β) Η πιθανότητα $\mathbb{P}[\ell(N) = k]$ μπορεί να υπολογιστεί με τη βοήθεια της συνάρτησης PRun παρατηρώντας ότι

$$\mathbb{P}[\ell(N) = k] = \mathbb{P}[\ell(N) < k + 1] - \mathbb{P}[\ell(N) < k].$$

Γράψτε μια συνάρτηση lrun στην R που θα δέχεται σαν είσοδο το N και θα επιστρέφει την κατανομή της τυχαίας μεταβλητής $\ell(N)$, δηλαδή ένα διάνυσμα διάστασης N , με συντεταγμένες $y_k = \mathbb{P}[\ell(N) = k]$ για $k = 1, 2, \dots, N$.

γ) Με ποιες εντολές στην R (χρησιμοποιώντας και τη συνάρτηση lrun του προηγούμενου ερωτήματος) θα υπολογίζατε τη μέση τιμή και τη διασπορά της τ.μ. $\ell(1000)$; Με ποια εντολή θα υπολογίζατε την τιμή του k για την οποία η πιθανότητα $\mathbb{P}[\ell(1000) = k]$ μεγιστοποιείται;

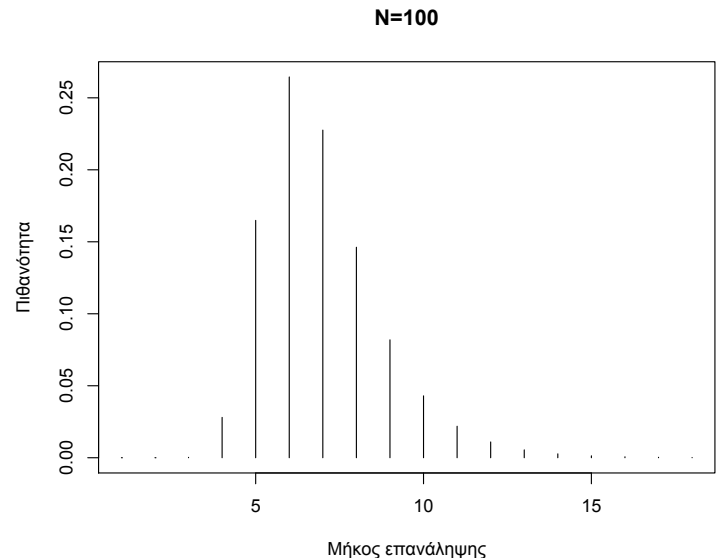
δ) Το διπλανό διάγραμμα παριστάνει την $\mathbb{P}[\ell(100) = k]$ για $k = 1, 2, \dots, 18$. Με ποιες εντολές της R θα μπορούσατε να το κατασκευάσετε;

Υπόμνηση: Αν δεν θυμάστε τη χρήση της $x[-1]$ που εμφανίζεται στην PRun θα βρείτε χρήσιμο το ακόλουθο παράδειγμα.

```

> z<-c(2,1,6)
> z[-2]
[1] 2 6

```



Άσκηση 2 (30 μονάδες)

Ένα τυχαίο δείγμα από 34 μαθητές της Β' Λυκείου (17 αγόρια και 17 κορίτσια) συμμετείχε σε ένα διαγώνισμα λεξιλογίου. Το εύρος του λεξιλογίου βαθμολογήθηκε στην κλίμακα 0 - 100 και οι βαθμοί των μαθητών φαίνονται παρακάτω.

Αγόρια	85	96	72	77	86	75	63	81	67	64	95	77	93	80	98	60	85
Κορίτσια	100	90	68	78	69	83	74	59	88	94	77	55	92	84	91	69	99

α) Ποιον γραφικό τρόπο θα επιλέγατε για να περιγράψετε τα παραπάνω αποτελέσματα σε ένα σχήμα, και με ποιες εντολές της R θα το υλοποιούσατε;

Με βάση το διαγώνισμα αυτό, θέλουμε να ελέγξουμε (σε επίπεδο σημαντικότητας 2%) την υπόθεση ότι το μέσο εύρος λεξιλογίου στους νέους αυτής της ηλικίας δεν διαφέρει ανάμεσα στα δύο φύλα.

Μια εντολή που παρεμβάλλεται στη στατιστική ανάλυση αυτού του ερωτήματος με τη βοήθεια της R είναι η

```
> var.test(male,female)
```

και τα βασικά αποτελέσματα από την εκτέλεση αυτής της εντολής είναι τα παρακάτω.

F=0.7845, num df=16, denom df=16, p-value=0.6332

β) Περιγράψτε τον έλεγχο που κάνει η παραπάνω εντολή, και σχολιάστε τα αποτελέσματα που δίνονται.

γ) Γράψτε τις εντολές της R που πιστεύετε ότι προηγήθηκαν της παραπάνω εντολής, και αναφέρετε για κάθε μια το σκοπό για τον οποίο εκτελεστήκε.

δ) Ποιες εντολές πιστεύετε ότι πρέπει να ακολουθήσουν την παραπάνω εντολή ώστε να ολοκληρωθεί σωστά η ανάλυση;

Άσκηση 3 (40 μονάδες)

Κατά τη διάρκεια του 2011 μελετήθηκε η συνολική κατανάλωση βενζίνης σε 48 πολιτείες των ΗΠΑ. Από κάθε πολιτεία συλλέχθηκαν δεδομένα για τις παρακάτω μεταβλητές.

- Consumption: Κατανάλωση βενζίνης (σε εκατομμύρια γαλόνια)
- Tax: Φόρος στη βενζίνη (σε σεντς/γαλόνι)
- Highways: Συνολικό μήκος αυτοκινητοδρόμων (σε μίλια)
- License: Ποσοστό κατόχων άδειας οδήγησης (%)

Θέλουμε να εξετάσουμε αν η κατανάλωση βενζίνης σε μια πολιτεία εξαρτάται από το φόρο στη βενζίνη, το συνολικό μήκος των αυτοκινητοδρόμων, και το ποσοστό κατόχων άδειας οδήγησης, προσαρμόζοντας το ακόλουθο γραμμικό μοντέλο.

$$\mathbb{E}[\text{Consumption} \mid \text{Tax, Highways, License}] = a + b_1 \text{Tax} + b_2 \text{Highways} + b_3 \text{License}.$$

α) Τα δεδομένα βρίσκονται σε μορφή πίνακα 48×4 στο αρχείο data.txt. Με ποιον τρόπο μπορείτε να τα εισάγετε στην R και να δημιουργήσετε ένα πλαίσιο δεδομένων, δίνοντας τα ονόματα που προαναφέρθηκαν στις 4 μεταβλητές;

β) Με ποιες εντολές της R θα προσαρμόζατε το παραπάνω γραμμικό μοντέλο;

γ) Ποιες προϋποθέσεις θα ελέγχατε και με ποιες εντολές της R;

δ) Εξηγήστε πλήρως τα ακόλουθα αποτελέσματα που πήραμε από την R ύστερα από την προσαρμογή του παραπάνω γραμμικού μοντέλου.

```
Residuals:
    Min       1Q   Median       3Q      Max
-142.16  -57.83    1.53    50.06   236.20

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  229.713004   208.378242    1.102  0.27629
Tax          -40.627490    14.783790   -2.748  0.00866**
Highways     -0.003950    0.003863   -1.023  0.31208
License       11.933177    2.165591    5.510  1.76e-06***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 76.09 on 44 degrees of freedom
Multiple R-squared:  0.567, Adjusted R-squared:  0.5374
F-statistic: 19.2 on 3 and 44 DF, p-value: 4.151e-08
```

ε) Εκτιμήστε την ετήσια κατανάλωση βενζίνης σε μια πολιτεία που διαθέτει αυτοκινητοδρόμους συνολικού μήκους 3111 μιλίων, το ποσοστό των κατοίκων με άδεια οδήγησης ανέρχεται στο 63%, και ο φόρος στη βενζίνη είναι 8 σεντς ανά γαλόνι.

στ) Με βάση το παραπάνω μοντέλο σε ποιο ποσοστό θα περίμενε αυτή η πολιτεία να αυξηθούν τα αναμενόμενα έσοδά της από το φόρο στη βενζίνη αν διπλασίαζε αυτόν το φόρο; Σχολιάστε το αποτέλεσμα.

Διάρκεια Εξέτασης 2 ώρες και 30 λεπτά
ΚΑΛΗ ΕΠΙΤΥΧΙΑ!