ΕΦΑΡΜΟΣΜΕΝΕΣ ΠΙΘΑΝΟΤΗΤΕΣ ΚΑΙ ΣΤΑΤΙΣΤΙΚΗ
Καθ. Γ. Κοντογιάννης
Τετάρτη 30 Οκτωβρίου 2013

Σετ ασκήσεων #2

ΕΝΤΟΛΕΣ: Για τις ασκήσεις αυτού του σετ, για τη δημιουργία τυχαίων αριθμών να χρησιμοποιηθεί **μόνο** η απλή εντολή `rand()`. Η εκθετική συνάρτηση e^x στο Scilab υπολογίζεται από την εντολή `exp(x)`.

1. *Ένας "randomized" αλγόριθμος.* Έστω μια συνάρτηση $F:\{0,1,\ldots n-1\}\to\{1,2,\ldots m\}$, τις οποίας τις τιμές τις έχουμε αποθηκευμένες σε έναν πίνακα, αλλά ξέρουμε πως το 1/5 των τιμών που περιέχει ο πίνακας είναι *λάθος*. Επίσης όμως ξέρουμε ότι η $F(x)$ έχει την εξής ιδιότητα:
$$F((x+y) \bmod n) = F(x)\,F(y) \text{ για κάθε } x \text{ και } y.$$
Περιγράψτε έναν απλό randomized αλγόριθμο ο οποίος να υπολογίζει την τιμή $F(x)$, και να μπορεί να μας εγγυηθεί πως η πιθανότητα αυτή η τιμή να είναι σωστή είναι τουλάχιστον 50%. Ο αλγόριθμός σας πρέπει να λειτουργεί μ' αυτόν τον τρόπο για *κάθε* x, και *χωρίς* να ξέρουμε ποιες από τις τιμές του πίνακα είναι λάθος.

2. *Πιθανοτική ανάλυση.* Στο μάθημα κάναμε πιθανοτική ανάλυση του κόστους εκτέλεσης C_N ενός αλγορίθμου που δρα σε μια ακολουθία N bits, $x_1, x_2, \ldots, x_N$, και το C_N είναι το μήκος της μακρύτερης ακολουθίας συνεχόμενων «1» ανάμεσα στα $x_1, x_2, \ldots, x_N$. Υποθέτοντας ότι τα bits αυτά προέρχονται από N ανεξάρτητες BERN(½) τυχαίες μεταβλητές $X_1, X_2, \ldots, X_N$, αποδείξαμε ότι η πιθανότητα το C_N να είναι μεγαλύτερο από $2\log_2 N$ τείνει στο μηδέν. Υποθέτοντας τώρα ότι τα $X_1, X_2, \ldots, X_N$ είναι N ανεξάρτητες BERN(p) τυχαίες μεταβλητές (για κάποιο p μεταξύ 0 και 1), αποδείξτε πως για οποιοδήποτε $\varepsilon > 0$, η πιθανότητα το C_N να είναι μεγαλύτερο από $k = \left\lceil \frac{1+\varepsilon}{\log_2(1/p)} \right\rceil \log_2 N$, τείνει στο μηδέν. Δικαιολογείστε όλα τα βήματα στον υπολογισμό σας. Πως μεταβάλλεται ο συντελεστής $\left\lceil \frac{1+\varepsilon}{\log_2(1/p)} \right\rceil$ για διαφορετικές τιμές του p; Σχολιάστε.

3. *Δημοσκόπηση*. Σε μια έρευνα αγοράς ερωτούνται N άτομα για το αν προτιμούν την WIND ή την COSMOTE, με σκοπό να εκτιμηθεί το (άγνωστο) ποσοστό των ατόμων που προτιμούν τη WIND.
   A. Πόσο μεγάλο πλήθος δειγμάτων N χρειαζόμαστε έτσι ώστε η απόκλιση της εκτίμησής μας από την πραγματική τιμή να είναι μικρότερη του 3% με πιθανότητα τουλάχιστον 90%?
   B. Πως θα άλλαζε η απάντηση αν γνωρίζαμε εκ των προτέρων ότι το ποσοστό που προσπαθούμε να εκτιμήσουμε είναι μικρότερο του 30%;

4. *Γεωμετρική κατανομή*.
   a. Περιγράψτε μια απλή μέθοδο για να προσομοιώσουμε τις τιμές μιας Τ.Μ. με ΓΕΩΜ(p) κατανομή, χρησιμοποιώντας μόνο ανεξάρτητες τυχαίες τιμές με κατανομή U[0,1], και με βάση την ερμηνεία της γεωμετρικής κατανομής ως χρόνου πρώτης «επιτυχίας» ενός πειράματος.
   b. Γράψτε ένα πρόγραμμα στο Scilab που να παράγει 100 ανεξάρτητα δείγματα από την ΓΕΩΜ(p) κατανομή.

5. *Μεταβλητές ελέγχου*. Έστω ότι έχουμε N ανεξάρτητα δείγματα $X_1, X_2, \ldots, X_N$ που το καθένα έχει ΓΕΩΜ(0.9) κατανομή, και θέλουμε να εκτιμήσουμε την μέση τιμή $\mu = E\left[e^{\sqrt{X_1}}\right]$. Δηλαδή την τιμή: $E\left[e^{\sqrt{X_1}}\right] = \sum_{k=1}^{\infty} e^{\sqrt{k}} \cdot (1-0.9)^{k-1} \cdot (0.9).$

   Προφανώς (γιατί?) είναι αδύνατο να κάνουμε αυτό τον υπολογισμό επακριβώς.
   α. Γεννήστε 1000 δείγματα $X_1, X_2, \ldots, X_{1000}$ όπως πιο πάνω, και κάντε ένα γράφημα των εκτιμήσεων σας για την μ ως συνάρτηση του N = 1,2,...,1000. Παρατηρήστε πως παρουσιάζουν σημαντικές διακυμάνσεις.
   β. Χρησιμοποιώντας το γεγονός ότι γνωρίζουμε πως η μέση τιμή των δειγμάτων $\nu = E(X_1)$ ισούται με $\nu = E(X_1) = 1/0.9 = 10/9 = 1.1111\ldots$, βελτιώστε την εκτίμησή σας με τη μέθοδο της «μεταβλητής ελέγχου» που είδαμε στο μάθημα. Δοκιμάστε διάφορες τιμές για την παράμετρο θ, και επιλέξτε μια με κριτήριο να μειώνει τις «διακυμάνσεις» των εκτιμήσεών σας. Προσπαθήστε να τη βρείτε με ακρίβεια τουλάχιστον ενός δεκαδικού ψηφίου. Ποια είναι η τιμή του θ που επιλέξατε;
   γ. Επαναλάβετε το β. αρκετές φορές (με την ίδια τιμή του θ που επιλέξατε), και δείξτε γραφικά τις δύο διαφορετικές εκτιμήσεις ως συνάρτηση του N. Ποια είναι η τελική σας εκτίμηση για το μ;
   δ. [**Προαιρετική**]. Βρείτε ένα τρόπο να *εκτιμήσετε* τη βέλτιστη τιμή του θ και ενσωματώστε τον στον πιο πάνω αλγόριθμο. Τι τιμές σας δίνει για το θ; Τι τιμές σας δίνει η νέα εκτιμήτρια (με τη μεταβλητή ελέγχου για αυτό το θ) για την εκτίμηση του μ;