האוניברסיטה העברית
המחלקה לסטטיסטיקה

מבחן סיום בקורס: **(52606) שיטות חישוביות בסטטיסטיקה גנטית**

תאריך הבחינה: 5.3.09, שעה 9:00. (מועד ב)
משך המבחן: שעתיים
חומר מותר בשימוש: 3 דפים כתובים בכתב ידו של הסטודנט.
הוראות: יש לענות על שתי השאלות ולנמק (בקיצור) כל אמירה. בהצלחה!

שאלה 1

יהי X משתנה מקרי המקבל ערכים שלמים בין 0 ל-10 ויהי Y משתנה מקרי המקבל ערכים ממשים בתחום $(0,1)$. נתון כי ההתפלגות של X, בהינתן Y, היא בינומית עם Y כהסתברות הצלחה. נניח כי ההתפלגות של Y היא $\text{beta}(5,5)$. כתוצאה מכך יתקבל כי ההתפלגות המותנית של Y, בהינתן X, היא $\text{beta}(5+X, 5+10-X)$. אנו מעוניינים לחשב את התוחלת של ההתפלגות השולית של X. נזין את הגדלים הרלוונטים:

```
> N <- 10
> alpha <- 5
> beta <- 5
```

להל"ן 2 פיסות קוד דומות, שמטרתן לחשב את התוחלת המבוקשת:

קוד א

```
> n.iter <- 10^2
> X <- 0
> X.sum <- 0
> for (i in 1:n.iter)
+ {
+   Y <- rbeta(1,alpha + X, beta + N-X)
+   X <- rbinom(1,N,Y)
+   X.sum <- X.sum + X
+ }
> X.mu <- X.sum/n.iter
> X.mu
[1] 4.75
```

קוד ב

```
> n.iter <- 10^2
> X <- 0
> X.sum <- 0
> for (i in 1:50)
+ {
+   Y <- rbeta(1,alpha + X, beta + N-X)
+   X <- rbinom(1,N,Y)
+ }
```

```
> for (i in 51:n.iter)
+ {
+    Y <- rbeta(1,alpha + X, beta + N-X)
+    X <- rbinom(1,N,Y)
+    X.sum <- X.sum + X
+ }
> X.mu <- X.sum/(n.iter-50)
> X.mu
[1] 4.94
```

1. מבין האפשרויות הבאות, מי המתאימה ביותר לאפיון האלגוריתם בו משתמשים?
   i. Bootstrap
   ii. EM
   iii. HMM
   iv. MCMC

2. הסבירו את ההבדל בין קוד א' לקוד ב'. באיזה קוד תעדיפו להשתמש? מדוע?
3. מה יקרה לדעתכם אם שורת הקוד "X <- 0" תוחלף בשורת הקוד "X <- 10"?
4. כתבו מחדש את קוד א' או את קוד ב' כך שתחושב, בנוסף לתוחלת, גם השונות השולית של X.

<u>שאלה 2</u>

חברה המעניקה שירותי אינטרנט מעוניינת לזהות את הלקוחות המשתמשים באתרים לשיתוף קבצים בכדי להפנות להם הצעות שיווקיות מתאימות. כלל זיהוי פשוט יכלול בקבוצה זו את כל המשתמשים בעלי נפח תקשורת מעל לסף מסוים. כדי לזהות סף מתאים נבחרה קבוצה אקראית של משתמשים בעבורם נקבע כתוצאה מבדיקה פרטנית האם הם משתמשים באתרים לשיתוף קבצים אם לאו. בנוסף, תועד נפח התקשורת של אנשים אלה. הסף נקבע להיות הסף שממזער את סך כל טעויות הסיווג, כאשר טעות סיווג כוללת משתפי קבצים שנפח התקשורת שלהם מתחת לסף ושאינם משתפים בעלי נפח תקשורת גבוה.

נסמן את נתוני האדם ה-i כ- (X_i, D_i) כאשר X_i הוא היקף התקשורת החודשית ו- D_i מקבל את הערך 1 אם האדם משתף קבצים ואת הערך 0 אחרת. נסמן ב- n את מספר האנשים שנמדדו.

1. נניח כי האחוז באוכלוסיה של המשתמשים בשירותי שיתוף קבצים הוא p והתפלגות נפח התקשורת שלהם מפולג $N(\mu_1, \sigma_1^2)$. נניח כי ההתפלגות בקרב הציבור שאינו משתמש בשירותים כנ"ל הוא $N(\mu_0, \sigma_0^2)$. רשמו ביטוי לתוחלת מספר הטעויות בעבור סף נתון x.
2. הציעו דרך, המבוססת על שיטת ה bootstrap הפרמטרי, כדי להעריך את תוחלת מספר הטעויות בעבור הסף שממזער את **מספר** הטעויות הנצפות.
3. חזרו על סעיף 2 בעבור bootstrap לא פרמטרי. איזו משתי השיטות תעדיפו?