

האוניברסיטה העברית
המחלקה לסטטיסטיקה

מבחן סיום בקורס: (52606) שיטות חישוביות בסטטיסטיקה גנטית

תאריך הבחינה: 22.2.09, שעה 12:00. (מועד א)

משך המבחן: שעתים

חומר מותר בשימוש: 3 דפים כתובים בכתב ידו של הסטודנט.

הוראות: יש לענות על שתי השאלות ולנמק (בקיצור) כל אמירה. בהצלחה!

שאלה 1

התפלגות קושי עם פרמטר מיקום θ נתונה בעזרת פונקציית הצפיפות $(1/\pi)(1+(x-\theta)^2)^{-1}$ על פני הישר הממשי. בהתפלגות זו התוחלת והשונות אינן מוגדרות. התפלגות זו משמשת לעתים קרובות כמודל לתצפית בה עשויות להתקבל ערכים חריגים. שימו לב כי ההתפלגות הינה סימטרית מסביב למדד המיקום ולכן, למרות שהתוחלת אינה מוגדרת, המדד שווה לחציון ההתפלגות.

נתון מדגם מגודל 30 מאוכלוסייה שהתפלגותה קושי בעלת פרמטר מיקום לא ידוע:

3.2, 1.65, 0.36, -29.92, 2.83, 35.1, 6.32, 1.67, 2.71, 1.33, 0.36, 2.96, 2.06, -6.17, 3.38, 2.00, 1.55, 2.27, 5.77, 3.41, 1.22, -73.48, 3.12, 3.08, 3.9, 1.37, 2.5, 0.35, 3.02, 1.94

התוצאות נשמרו ב-R באובייקט בשם X. נחשב אומד לפרמטר המיקום של האוכלוסייה ולתחום הבין-רבעוני שלה:

```
> hat.theta <- median(X)
> hat.theta
[1] 2.168676
> Q <- quantile(X,c(0.25,0.75))
> Q
  25%    75%
1.338195 3.107623
> Q[2]-Q[1]
[1] 1.769428
```

מעוניינים לאמוד את התחום הבין-רבעוני של האומד לפרמטר המיקום. לצורך כך הופעל הקוד שלהלן:

```
> n <- 10^4
> X.sim <- matrix(rcauchy(30*n, location = hat.theta),nrow=n,ncol=30)
> theta.sim <- apply(X.sim,1,median)
> Q.sim <- quantile(theta.sim,c(0.25,0.75))
> Q.sim[2]-Q.sim[1]
[1] 0.3868865
```

1. הסבירו את ההבדל שבין ההפרש $Q[2]-Q[1]$ להפרש $Q.sim[2]-Q.sim[1]$.
2. כיצד להערכתכם יושפע ההבדל מהסעיף הקודם אם גודל המדגם ישתנה מ-30 ל-300? אם ערכו של n יוחלף מ-10,000 ל-100,000?
3. נטען כי האומד לתחום הבין-רבעוני של האומד לפרמטר המיקום מבוסס על שיטת ה-bootstrap. האם אתם מסכימים עם טענה זו? אם כן, האם המדובר בגישה הפרמטרית או בגישה שאיננה פרמטרית?
4. ניתן להראות כי לאומד לפרמטר המיקום יש שונות סופית, למרות שלהתפלגות כל אחת מן התצפיות השונות הינה אין-סופית. שנו את הקוד שהופעל לאמידת התחום הבין-רבעוני של האומד כך שיתקבל אומד לסטיית התקן של האומד לפרמטר המיקום.

שאלה 2

מערך ניסוי בוחן רגישותם של תאי שמר לתרכובת נתונה. במסגרת הניסוי חושפים מספר נתון, נאמר n , של מושבות תאים לתרכובת וסופרים את מספר המושבות ששרדו בתום תקופת הניסוי. על הניסוי חוזרים פעמיים בתנאים זהים. יהיו X ו- Y מספר המושבות ששרדו בחזרה הראשונה והשנייה, בהתאמה. מעוניינים בחישוב מקדם המתאם בין שני המשתנים הנ"ל.

מודל הסתברותי למערכת מניח התפלגות א-פריורי ביתא להסתברות ההישרדות של מושבה אקראית. בהינתן הסתברות זו, הישרדותה או אי-הישרדותה של כל מושבה בלתי תלויה ברעותה. כפועל יוצא מתקבל כי ההתפלגות המותנית של X ושל Y , בהינתן ההסתברות להישרדות p , היא $B(n, p)$ ושני המשתנים בלתי תלויים זה בזה.

1. רשמו את ההתפלגות המשותפת של X ו- Y , כלומר את ההסתברות $P(X=x, Y=y)$, ונמקו מדוע שני המשתנים אינם בלתי תלויים. זכרו כי הצפיפות של התפלגות ביתא נתונה על ידי $p^{\alpha-1}(1-p)^{\beta-1}/B(\alpha, \beta)$ בעבור $0 < p < 1$ ובעבור $\alpha > 0$ ו- $\beta > 0$ פרמטרים נתונים.
2. רשמו את ההתפלגות המשותפת של X , Y ו- p והראו כי ההתפלגות המותנית של p , בהינתן X ו- Y , היא ביתא עם הפרמטרים $\alpha + X + Y$ ו- $\beta + 2n - X - Y$, בהתאמה.
3. הציעו אלגוריתם, המבוסס על שיטתו של Gibbs, המאפשר יצירה של סידרה של זוגות ערכים $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ כך שהתפלגות כל זוג ערכים הינה בקירוב ההתפלגות שתיארתם בסעיף 1.
4. כיצד ניתן להשתמש בסדרת הזוגות הנ"ל כדי לחשב את מקדם המתאם המבוקש?