מעבדה לסטטיסטיקה 2024, מטלה 7
הצגה ב9.7

חלק מקדים

אמנו מודל רגרסיה עבור תאים בגודל 2500 בסיסים, הן לדגימת הסרטן והן לדגימה הבריאה:

a. קבוצות A יעבדו על דגימת סרטן: TCGA-13-0723-01A_lib2_all-chr1
ועל דגימה בריאה: TCGA-13-0723-10B_lib2_all-chr1
b. קבוצות B,C יעבדו על דגימת סרטן: TCGA-13-0723-01A_lib1_all-chr1
ועל דגימה בריאה: TCGA-13-0723-10B_lib1_all-chr1

לצורך כך יש להכין מערך של אחוז GC בגודל התא המתאים.
כמו כן, לכל אחת משתי הדגימות, יש לחשב כיסוי עבור תאים בגודל 2500 בסיסים; לסמן ולהוציא ערכים יוצאי דופן או שאינם מועילים לרגרסיה; להחליט על בסיס spline מתאים; ולאמוד את פונקציית הרגרסיה.

אני אתן לכם גם קבצים באגרגציה ל100 בסיסים כפי שהיו בקוד Lecture7_new ואפשר להשתמש בהם.

חלק א׳ פירוק שונות (דגימה בריאה בלבד)

א. הציגו את פירוק השונות של הנתונים, אחרי סינון הערכים יוצאי הדופן ואחרי ״תיקון החציון״.
איזה אחוז מהשונות בדגימה הבריאה נובעת מהתפלגות הפוואסון?
איזה אחוז מהשונות בדגימה נובעת מאפקט ה GC ?
איזה אחוז מהשונות לא מוסברת על ידי גורמים אלו?

חלק ב׳ השוואת דגימות

ב. השוו את פונקציות הרגרסיה המתקבלות עבור שתי הדגימות. האם יש קשר דומה בין כיסוי לGC? האם הקשר זהה?

ג. אחרי שאמדתם את הרגרסיה על כל התאים, ננסה להשוות את התוצאות באזור בין הבסיס ה25 מיליון וה30 מיליון בלבד. (שימו לב איפה אזור זה נופל בתאים שלכם)

a. אמדו את מספר העותקים בכל דגימה.
הציגו את התוצאות לאורך הכרומוזום. השוו לאומדים שמתקבלים בעזרת תיקון חציון בלבד. באיזו מידה תיקון המבוסס על GC מועיל בכל אחת מהדגימות.

b. הסתכלו בדגימה הסרטנית. זהו רצפי תאים בהם ייתכן שיש שינוי במספר עותקים.

c. כעת השוו לדגימה הבריאה. אילו מהרצפים שזיהיתם דומים בשתי הדגימות?
האם יש רצף תאים שמראה הבדל בין הדגימה הסרטנית לבין הדגימה הבריאה?

d. צרו גרף פיזור המשווה בין מספר העותקים בתא של הדגימה הבריאה (ציר הx) ושל הדגימה הסרטנית (ציר הy). מה רואים? הציעו רעיונות כיצד להשתמש בדגימה

הבריאה כדי לקבל אומדנים טובים יותר עבור הדגימה הסרטנית (אין צורך לבצע רעיונות).