מעבדה לסטטיסטיקה: אמידת מספר עותקים מדגימה בודדת (א)

אנחנו רוצים להשתמש בפונקציית ה GC שלמדנו ובערכים הנצפים Y_k כדי לאמוד את מספר-העותקים (Copy number) של הגנום.
בסיכום הזה נציג שיטת אמידה שדורשת רק אמידה של פונקציית ההשפעה של GC. לעיתים, קוראים לשיטות אמידה אלו "תיקון", משום שהן מתקנות את ההטיה הנוצרת מאפקט הGC.

מודל המתאר קשר בין מספר עותקים למספר פרגמנטים

נשתמש ב k=1…K לתאר את מספר התא, וב Y_k לתאר את הכיסוי הנצפה בתא.
נסמן את התוחלת של התא הk בλ_k

$$\lambda_k = E[Y_k].$$

המודל עבור התוחלת צריך לסמן את הרכיבים שמרכיבים את התוחלת, ולאפשר לנו לתאר איך היינו מייצרים נתונים כאלו. נשים לב ל3 רכיבים שאנחנו יודעים שמשפיעים על הכיסוי:

- כמות העותקים הוא מספר בכפולות שלמות של חצי (…0, ½, 1, 1 ½ , 2). תוחלת הכיסוי צריכה להיות פרופורציונלית למספר העותקים.
- סה"כ כמות הפרגמנטים (ע"פ כל התאים) יכולה להשתנות מניסוי לניסוי, ואינה קשורה למספר העותקים אלא למכונות הריצוף, נסמנה בN.
- לכמות הGC בתא (או נגזרות כמו כמות הGC בתתי התאים) יש השפעה על הכיסוי. תיארנו תלות זו בעזרת מודלים של רגרסיה.

אם כן, נוח לתאר את התוחלת כפונקציה כפלית של שלושה גורמים <u>בלתי תלויים,</u>
כפול גורם לא ידוע η_k:

$$\lambda_k = N \cdot a_k \cdot f(gc_k) \cdot \eta_k,$$

לצורך זיהוי הפרמטרים צריך להניח הנחות לגבי הממוצעים שלη_k, $f(gc_k)$ ושל a_k. נבחר לדרוש מהחציון של a_k להיות 1, והממוצעים של הגורמים האחרים להיות 1:

$$med_k\{a_k\} = 1, \quad avg_k\{f(seq_k)\} = 1, \quad avg_k\{\eta_k\} = 1,$$

ואז N הקבוע שמתקן בהתאם, כלומר

$$N \approx med_k\{\lambda_k\}$$

אמידת מספר עותקים מהמודל

כעת נניח שיש לנו אומדן לפונקציה $\hat{f}(seq_k)$ ואת מספרי הפרגמנטים הנצפים $y_1, \dots, y_k$, ואנחנו רוצים לאמוד את מספר העותקים a_k.
(הפונקציה $\hat{f}(gc_k)$ כפי שאמדנו בכיתה לא תקיים $avg_k\{\hat{f}(gc_k)\} = 1$ אלא $avg_k\{\hat{f}(seq_k)\} = avg_k\{\lambda_k\}$ ולכן נצטרך לשים לב לקבועים בהמשך.)

ראשית נחלץ את a_k מהמודל:

$$a_k = \frac{\lambda_k}{f(gc_k)} \cdot \frac{1}{N} \cdot \frac{1}{\eta_k},$$

וכעת נשתמש באומד ההצבה כאשר נציב y_k במקום λ_k ואת $\hat{f}(gc)$ במקום $f(gc_k) \cdot N$.

$$\widetilde{a_k} = \frac{y_k}{\hat{f}(gc_k)}.$$

הערות

1. ניתן להכפיל בעוד קבוע M[1] שידאג שהחציון של האומדים יהיה 1:
$$\hat{a}_k = \frac{y_k}{\hat{f}(gc_k)} \cdot \hat{M}, \qquad \hat{M} = \frac{1}{med_k\{\widetilde{a_k}\}}.$$
2. ייתכן שיש עוד גורמים שתלויים ברצף הגנומי שמשפיעים על λ_k, כמו לדוגמה הסיכוי למפות פרגמנט, ואולי דרכי מדידה יותר עדינות ממספר הgc בתא. כדי לסמן מצב כזה, ניתן להשתמש ב $f(seq_k)$ בתיאור המודל במקום $f(gc_k)$. ככל שנכניס יותר גורמים ל$f(seq_k)$, כך המשמעות של η_k תשתנה ובתקווה השונות שלו תקטן.

---

[1] M מגלם גם את 1/N וגם את הממוצע של $1/\hat{f}(seq_k)$

מעבדה לסטטיסטיקה: אמידת מספר עותקים (2)

אנחנו רוצים להשתמש בפונקציית ה GC שלמדנו ובערכים הנצפים Y_k לבין מספר-העותק של הגנום. ראשית נדבר על אמידה שאיננה דורשת דוגמת control, ואח״כ נציג אמידה שמשתמשת בדוגמת control. לעיתים, קוראים לשיטות אמידה אלו ״תיקון״, משום שהן מתקנות את ההטיה הנוצרת מאפקט הGC.

1. אמידה לתא בדגימה בודדת

1.1.מודל

נניח שמספר הפרגמנטים Y_k הממופים לתא הk מתפלגים $Poisson(\lambda_k)$. אנחנו נבנה מודל לפרמטר התוחלת λ_k שנותן שמות לתרומה של הGC ומספר העותקים.

נסמן ב $a_k \in \{0, \frac{1}{2}, 1, \frac{3}{2}, 2 \ldots\}$ את מספר העותקים, כאשר בדרך כלל $c_k = 1$. נניח ששינוי מספר העותקים משפיע כפלית על מספר הפרגמנטים הנצפים. נגדיר לכן:

$$\lambda_k = a_k \cdot \mu_k$$

כאשר μ_k הוא הערך הצפוי של פרגמנטים אם היה עותק יחיד. נזכור, שאנחנו ניסינו להבין כיצד μ_k מושפע מהרצף הגנומי המרכיב את התא הk , לדוגמה מכמות הGC בתא או מתוך כמות הGC באזורים בתא. נסמן פונקציה זו ב $f(GC_k)$. ראינו שהפונקציות אכן מנבאות חלק אך לא את כל התוחלת של התא. אם כן, ניתן לכתוב:

$$\lambda_k = a_k \cdot f(GC_k) \cdot \beta_k, \qquad Y_k = c_k \cdot f(GC_k) \cdot \beta_k \cdot \eta_{k,Pois}$$

כאשר β_k היא השגיאה בתוחלת של התא שאיננה קשורה בGC, ו $\eta_{k,Pois}$ מסמן את השגיאה הנובעת מההגרלה הפוואסונית. נניח ש $E_k[\beta] = 1$ וש $E_{Pois}[\eta_{k,Pois}] = 1$.

הערות:

א. לשם נוחות ההצגה השמטתי את הקבוע J המסמן את התוחלת הכללית. מבחינת הניסוי, בד״כ כן מתעניינים בקבוע זה כדבר נפרד מפונקציית $f(GC_k)$.

1.2.אמידה בדגימה בודדת

כעת, נניח שאמדנו את הפונקציה $f(GC_k)$ בעזרת רגרסיה, ונסמן את הפונקציה שנאמדה ב $\hat{f}(GC_k)$. אם כן, ניתן לאמוד את c_k ע״י הצבה (plug-in) במשוואת התוחלת:

$$E_k[E_{Pois}[Y_k]] = c_k \cdot f(GC_k)$$

$$a_k = \frac{E[Y_k]}{f(GC_k)}$$

ולכן, מעקרון אמידת מומנטים נקבל את האומד הבא:

$$\hat{a}_k^{(1)} := \frac{Y_k}{\hat{f}(GC_k)}.$$

1.3.התפלגות ע״פ המודל

נציב את המודל Y_k, ונקבל:

$$\hat{a}_k^{(1)} = \frac{a_k \cdot f(GC_k) \cdot \beta_k \cdot \eta_{k,Pois}}{\hat{f}(GC_k)} = a_k \cdot \frac{f(GC_k)}{\hat{f}(GC_k)} \cdot \beta_k \cdot \eta_{k,Pois}.$$

נסמן את השונות של $\beta_k \cdot \eta_{k,Pois}$ ב σ^2, ונניח שהאמידה של f טובה ולכן $\frac{f(GC_k)}{\hat{f}(GC_k)} \approx 1$. אם כן ההטיה של $\widehat{a_k}^{(1)}$ קרובה ל-0, והשונות של $\widehat{a_k}^{(1)}$ בקירוב $c_k^2\sigma^2/\hat{f}^2(GC_k)$:

$$Bias[\hat{a}_k^{(1)}] =$$

$$E_k E_{Pois}\left[c_k \cdot \frac{f(GC_k)}{\hat{f}(GC_k)} \cdot \beta_k \cdot \eta_{k,Pois}. - a_k\right] = a_k\left(\frac{f(GC_k)}{\hat{f}(GC_k)} \cdot E_k E_{Pois}[\beta_k \cdot \eta_{k,Pois}] - 1\right) \approx 0.$$

$$Var[\hat{a}_k^{(1)}] = \frac{Var[Y_k]}{\hat{f}(GC_k)^2}.$$

1.4.ייצוב האומד (בעזרת epsilon)

בעזרת תוספת של $\epsilon > 0$ הן למונה והן למכנה, ניתן לשנות את האיזון בין הטיה לשונות. נגדיר את האומד

$$\hat{a}_k^{(1*)} = \frac{Y_k + \epsilon}{\hat{f}(GC_k) + \epsilon}.$$

מבחינת ההתפלגות, נשים לב שעבור $a_k \neq 1$, ההטיה של האומד $\hat{a}_k^{(1*)}$ עם הייצוב גדלה.

מצד שני, השונות של $\hat{a}_k^{(1*)}$ היא בקירוב $\frac{a_k^2\sigma^2}{(\hat{f}(GC_k)+\epsilon)^2}$ ועל כן קטנה יותר מהשונות של $\hat{a}_k^{(1)}$.

2. **אמידה בשתי דגימות עם GC**

כעת, נדבר על אמידה (או תיקון) בעזרת דגימה נוספת. במקרה כזה, נגדיר את מספר העותקים של הדגימות הבריאה להיות תמיד 1. כמו כן, נפרק את השונות של תוחלת-התא לרכיב שמשותף לשתי הדגימות, ולרכיב ששונה בין הדגימות.

נסמן ב Y_k^t את דגימות הסרטן וב Y_k^n את הדגימות הנורמליות. נניח את המודל הבא:

$$Y_k^t = a_k \cdot f^t(GC_k) \cdot \gamma_k \cdot \delta_k^t \cdot \eta_{k,Pois}^t,$$

$$Y_k^n = f^n(GC_k) \cdot \gamma_k \cdot \delta_k^n \cdot \eta_{k,Pois}^n.$$

פיצלנו את β_k^t, β_k^n , לרכיב המשותף γ_k בין הדגימות ולרכיבים הנפרדים β_k^n , δ_k^t ,.

האומד של מספר העותקים המבוסס על שתי דגימות יהיה בצורה הבאה:

$$\hat{a}_k^{(2)} = \frac{Y_k^t/\hat{f}^t(GC_k) + \epsilon}{Y_k^n/\hat{f}^n(GC_k) + \epsilon}$$

2.1. התפלגות ע״פ המודל

עם נתעלם רגע מהepsilon

$$\hat{a}_k^{(2)} = a_k \cdot \frac{f^t(GC_k)}{\hat{f}^t(GC_k)} \frac{f^n(GC_k)}{\hat{f}^n(GC_k)} \cdot \frac{\delta_k^t \cdot \eta_{k,Pois}^t}{\delta_k^n \cdot \eta_{k,Pois}^n} \approx a_k \cdot \frac{\delta_k^t \cdot \eta_{k,Pois}^t}{\delta_k^n \cdot \eta_{k,Pois}^n}.$$

אם כן, את הרעש ב אומד דגימה אחת $\hat{a}_k^{(1)}$ שהיה $\beta_k^t \cdot \eta_{k,Pois}^t = \gamma_k \cdot \delta_k^t \cdot \eta_{k,Pois}^t$, אנחנו מחליפים ב $\frac{\delta_k^t \cdot \eta_{k,Pois}^t}{\delta_k^n \cdot \eta_{k,Pois}^n}$.

3. תיקוני חציון

אם אנחנו מודדים את איכות האומדים על ידי השוואה לערך המקובל 1, כדאי קודם כל לדאוג שהנתונים אכן מיושרים ל1. לצורך כך, ניתן לחלק כל אחד מוקטורי האומדים בחציון שלהם, וכך נבטיח שחציון האומד יהיה 1.