**תרגול 9 – FDR ומרחקים**

בשבוע שעבר התוודענו לבעיית ההשוואות המרובות אשר מעלות את הסיכוי לטעויות מסוג 1. ראינו את תיקון בונפרוני שעוזר לתקן את הבעיה, אך הוא מחמיר, מקשה על מציאת הבדלים אמיתיים בין קבוצות ומעלה את הסיכוי לטעות סוג 2.

היום נפגוש גישה חילופית המאזנת בין הצורך להגיע למסקנות במחקר לבין הרצון לצמצם את מספר הטעויות מסוג ראשון.

**FDR (False Discovery Rate)**

זוהי גישה להתמודדות עם בעיית ההשוואות המרובות על ידי שערוך סך הטעויות מסוג 1 **בכל ההשוואות**, כפונקציה של רמת המובהקות שדורשים בהשוואה בודדת. אחרי השערוך קל לנו יותר להחליט מהי רמת המובהקות הנדרשת על מנת לקבל מספר "נסבל" של שגיאות.

$$FDR=\frac{False Positives}{True Positives+False Positives}$$

אם מדברים על בדיקה רפואי למשל, ה-FDR הוא היחס של מספר הבדיקות שיצאו false positive מתוך כלל התוצאות שיצאו חיוביות. במבחן סטטיסטי, אם יש FDR=0.05 אזי 5% מכל המדדים שיצאו מובהקים, הם בעצם לא.

ישנן מספר שיטות להעריך את ה-FDR – ואנו נשתמש בפרמוטציות שפגשנו בשיעור הקודם. כלומר, להעריך את כמות השגיאות באמצעות הנתונים הגולמיים על ידי ערבול הקבוצות. בעוד שבשיעור הקודם השתמשנו בפרמוטציות כדי להעריך מובהקות של השוואה (מבחן) בודדת, הפעם ניישם את הפרמוטציות עבור כל ההשוואות.

אם מערבבים את שתי הקבוצות כך שיש לנו שתי קבוצות אקראיות, אנחנו לא מצפים שיהיו הבדלים אמיתיים ביניהם, וכל הבדל "מובהק" שיימצא הוא בבירור שגיאה מסוג 1. אם נספור בכמה מימדים (מדדים) נמצאו הבדלים "מובהקים" בין הקבוצות המעורבבות, ונשווה לכמות המדדים שבהם יש הבדל מובהק בקבוצות המקוריות, נוכל להעריך כמה מתוך הבדלים בין הקבוצות האמיתיות הם למעשה טעות מסוג 1. את כל זה נעשה תחת ערך אלפא מסויים.

איך עושים את זה בפועל?

1. מוצאים p-value לכל אחת מההשוואות (בחינות, בדוגמא של הריטלין)
2. מערבבים את הקבוצות ושוב מוצאים p-value לכל אחת מההשוואות
3. חוזרים על סעיף 2 הרבה פעמים
4. עבור רמת המובהקות שאנחנו רוצים, נבדוק כמה השוואות יצאו מובהקות הן בקבוצות המקוריות (בסעיף 1) והן בקבוצות המעורבבות (בסעיף 2).
5. מחשבים FDR:

$$FDR\_{permutation}=\frac{mean \left(\# of significant comparisons in shuffled group\right)}{\# significant comparisons in original group}$$

המשמעות של התוצאה: FDR הוא הסיכוי לטעות מסוג ראשון **בכלל ההשוואות**, בהינתן אלפא מסויים **בהשוואה בודדת**. ככל שמספר הפרמוטציות יהיה גדול יותר, נקבל דיוק גבוה יותר בחישוב ה-FDR.

שימו לב: יש לזכור לבצע פרמוטציות על המשתנה הידוע. כלומר אם יודעים מי לוקח ריטלין ומי לא, מתייגים כעת באקראי עם/בלי ריטלין. לא מערבבים את הבחינות (ה"מימדים"), אלא את התיוג של הסטודנטים (ה"דוגמאות").

על מנת לבחור $α$ נרצה לדעת מה ה-trade off שלנו בין מספר השוואות מובהקות לבין הטעות הסטטיסטית. נסתכל על דוגמת הריטלין מהשבוע שעבר ונעשה עליה 100 פרמוטציות לכל מבחן.

נשתמש בפונקציה sample:

p = sample(n)

n - אורך הוקטור עליו אנו עושים את הפרמוטציה

p – וקטור אינדקסים במיקומים אקראיים באורך n.

load("ritalin\_data.Rdata")

alpha<-0.05

#define the groups to be compared

group1 <- 1:39

group2 <- 40:78

#find p-values for each comparison

data\_p <- numeric(dim(Data)[1])

for (i in 1:40) {

 data\_p[i]<-t.test(Data[i,group1],Data[i,group2])$p.value

}

#find p-values for each comparison in shuffled data

number\_of\_permutations<-100

shuffled\_p <-matrix(0,dim(Data)[1],number\_of\_permutations)

for (j in 1:number\_of\_permutations) {

#shuffle groups

permute<-sample(78)

group1\_permutated<- permute[1:39]

group2\_permutated<- permute[40:78]

#find p-values for the shuffled groups

for (i in 1:40) {

shuffled\_p[i,j]<-t.test(Data[i,group1\_permutated],Data[i,group2\_permutated])$p.value

}

}

alpha<-seq(0.005,0.3,0.005)

FDR<-numeric(length(alpha))

data\_significance<-numeric(length(alpha))

#compute FDR for each alpha

for (k in 1:length(alpha)) {

data\_significance[k]<- sum(data\_p < alpha[k])

shuffled\_significance <- colSums(shuffled\_p < alpha[k])

FDR[k]<-mean(shuffled\_significance)/data\_significance[k]

FDR[k]<-max(FDR[1:k]) # avoid non-monotonic portions

}

#plot

plot(data\_significance,FDR,type='b',xlab='# significant comparisons',ylab='FDR')



זהו הגרף שאומר לנו מהו ה"מחיר" (מבחינת טעויות מסוג 1) של העלאת מספר המובהקויות. נשים לב שהגרף בנוי מ"מדרגות" – כלומר יש מספר נקודות בהם נוכל להעלות את מספר המובהקויות מבלי לשלם "מחיר". אם מטרת הניסוי שלנו היא למקסם את מספר המובהקויות שנקבל, נבחר $α$ בדיוק לפני עליית מדרגה. בנוסף ניתן לראות שיש מקומות שיש בהם קפיצות משמעותיות (למשל ב-17 השוואות קפצנו מ-$α=0.25$ ל-$α=0.36$).

למה זה קורה? כי ככל ש-$α$ עולה נגלה יותר מובהקויות אמיתיות (שקודם היו $p>α$) אבל מצד שני נקבל גם יותר טעויות מסוג 1. החלק הראשון קובע את רוחב המדרגה (מספר ההשוואות המובהקות) והחלק השני את גובה המדרגה (ה-FDR).

*איך קובעים את אלפא האופטימלי? זה כבר תלוי מה רוצים לעשות עם התוצאות, ועד כמה אנחנו מוכנים "לסבול" טעויות מסוג 1, לשיקול דעתו הסובייקטיבי של עורך הניסוי.*

***דמיון ומרחק***

*דמיון הוא מדד כמותי לכמה אובייקטים מרוחקים אחד מהשני – ככל שהם יותר רחוקים, כך הם פחות דומים. כאשר יש לנו מידע חד-מימדי, למשל רמות סוכר בדם, קל לנו להשוות בין רמות הסוכר בדם בין שני מטופלים ולהגיד אם הם דומים או שונים זה מזה. אך ככל שמספר המדדים שלנו גדל, כך קשה לנו יותר (אינטואיטיבית) להגדיר מרחק ודמיון בין נקודות.*

*נסתכל על הציונים של 4 סטודנטים ב-30 בחינות וננסה להבין איזה סטודנט הכי "דומה" לסטודנט 1:*

load("grades.Rdata")

matplot(t(grades), type = c("b"),pch=1,xlab='Exam',ylab='Grades',lwd = 2)

legend("topright", legend = c("Student 1","Student 2","Student 3","Student 4"),col=1:4,pch=1)



לכאורה נראה שסטודנט 2 הכי דומה לסטודנט 1. אבל זה תלוי איך מגדירים מרחק. מיד נראה שהתשובה האינטואיטיבית שלנו מתאימה להגדרה של מרחק אוקלידי.

מרחק אוקלידי

בשני מימדים, מרחק אוקלידי מוגדר על ידי משפט פיתגורס:

$$d=\sqrt{\left(x\_{2}-x\_{1}\right)^{2}+\left(y\_{2}-y\_{1}\right)^{2}}$$

ברב מימד המרחק מוגדר באופן דומה:

$$d\left(x\_{1},x\_{2}\right)=\sqrt{Σ\_{i}\left(x\_{2}^{i}-x\_{1}^{i}\right)^{2}} $$

כעת נחזור לנתונים שלנו, ונמצא את המרחק בין כל סטודנט לסטודנט לפי הנוסחא הזו. נקבל מטריצה של 4X4 (כי יש 4 סטודנטים, ועבור כל סטודנט אנו בודקים את המרחק משלושת הסטודנטים האחרים). המטריצה הזו נקראת **מטריצת המרחקים (distance matrix)**.

נקבל טבלת מרחקים:

Euc\_distance <- round(as.matrix(dist(grades,method="euclidean")),2)



אותה אפשר לקרוא, אבל יותר קל להסתכל עליה עם heatmap:

BiocManager::install("ComplexHeatmap")

Heatmap(Euc\_distance, cluster\_columns = F, cluster\_rows = F, name = "Euclidean Distance")



נשים לב למספר דברים:

ראשית, יש לנו קו אלכסוני בצבע כחול (= מרחק 0 לפי סקלת הצבעים) שנמצאת בדיוק בנקודת המפגש של כל סטודנט עם עצמו. זה הגיוני מכיוון שאין מרחק בין כל סטודנט לעצמו.

שנית, שימו לב שהתמונה משני צידי האלכסון היא זהה – כלומר, מספיק לנו לדעת רק את אחד הצדדים של המטריצה כדי לדעת את המרחקים.

לבסוף – על מנת למצוא את המרחק המקסימלי נסתכל בגרף ונראה שהמרחק הגדול ביותר הוא בין סטודנט 2 לסטודנט 4 (הצבע הכי צהוב), ואכן כשנסתכל בטבלה נראה שיש לו את הערך המקסימלי 289.6. כאשר נחפש את המרחק הקטן ביותר מסטודנט 1 נראה שסטודנט 2 הכי רחוק ממנו עם הערך 73.3.

כעת נבחן הגדרות מרחק אחרות שיובילו לתשובה אחרת.

קורולציית פירסון (Pearson)

קורולציית פירסון מודדת איך שני משתנים משתנים יחד. במילים אחרות, קורולציית פירסון מודדת את ה-covariance של שני משתנים (co – יחד, variance – שונות 🡨 משתנים יחד) ומחלקת אותו בסטיות התקן של המשתנים. לדוגמא, ייתכן שממוצעי הציונים של מיטל ואריק שונים לחלוטין, אבם אם הם טובים באותם מקצועות וגרועים באותם מקצועות (כל אחד ביחס לממוצע שלו) – אז יש בינהם קורולציה חיובית.

$$r\_{x,y}=\frac{1}{n}⋅\frac{\sum\_{i}^{n}\left(\left(x\_{i}-\overbar{x}\right)⋅\left(y\_{i}-\overbar{y}\right)\right)}{σ\left(x\right)⋅σ\left(y\right)}=\frac{covariance\left(x,y\right)}{σ\left(x\right)⋅σ\left(y\right)}$$

כאשר $r\_{x,y}$ הוא מקדם פירסון, $σ$ הוא סטיית התקן ו-$n$ מספר המימדים.

נשים לב שעבור כל מימד נקבל ביטוי חיובי או שלילי לפי כיווני הסטיות של $x\_{i}$ ו- $y\_{i}$ מהממוצעים שלהם (חיובי אם שניהם סוטים מהממוצעים שלהם באותו כיוון, ושלילי אם אחד מעל הממוצע שלו והשני מתחת לממוצע שלו).

הערכים של קורולציית פירסון תמיד יהיו בין -1 ל-1 כאשר אלו הם הערכים אשר מבטאים קורולציה מלאה (חיובית או שלילית), ו-0 מבטא חוסר קורולציה מוחלט.

כדי למצוא את המרחק בעזרת קורולציית פירסון נחשב: $d=1-r$.

הערה: אפשר לחשב את המרחק גם לפי $d=1-|r|$. דרך החישוב תלוייה בהקשר. אם אנחנו מעוניינים לבדוק אם יש קורילצייה כלשהי בין אם קורילציה חיובית או קורילציה שלילית נוכל להשתמש ב $d=1-|r|$ . אם חשובה לנו מגמת הקורילציה נשתמש ב $d=1-r$.

נחזור לשאלה איזה סטודנט הכי דומה לסטודנט הראשון נחשב את המרחק d בעזרת קורולציית פירסון:

pearson\_distance<-round(1-cor(t(grades),method="pearson"),2)



*שימו לב בחישוב המרחק כאן ה-*$r$ *לא בערך מוחלט.*

*ניתן לראות כי לפי קורולציית פירסון הסטודנט שהכי קרוב לסטודנט 1 הוא סטודנט 3 עם מרחק של 0.06. איך לא ראינו את זה קודם בגרף? אם נסתכל שוב על הגרף הראשון נראה שההתנהגות של שני העקומים נראית כמעט זהה – ואכן, כאשר מחשבים מרחק אוקלידי שתי דוגמאות קרובות יהיו גם קרובות בגרף – אבל כאשר מחשבים פירסון, שתי דוגמאות קרובות יראו דומות, למרות מרחק אוקלידי שיכול להיות גדול.*

Heatmap(pearson\_distance, cluster\_columns = F, cluster\_rows = F, name = "Pearson\nDistance", col = c("#fff9e3","#719d92","#003f5c"))



*דרך נוספת להציג את הקורולציה בין שתי דוגמאות היא להציג בגרף את הערכים של דוגמא אחת נגד הערכים של דוגמא אחרת. ככל שהיחס בין הדוגמאות יותר לינארי, הקורולציה תהיה גבוהה יותר. הגרף הבא מציג את הקורולציה של סטודנט 1 מול 3 הסטודנטים האחרים:*

plot(grades[1,],grades[2,],type='p',xlim=range(grades[1,]),ylim=range(grades[2:4,]), pch=16,xlab="Student 1",ylab="Other students")

lines(grades[1,],grades[3,],type='p', col='blue',pch=16)

lines(grades[1,],grades[4,], type = 'p',col='red',pch=16)

legend("topright", legend = c("Student 2","Student 3","Student 4"),

 col=c("black","blue","red"),pch=16)



*אכן ניתן לראות שהקורולציה עם סטודנט 3 נראית הכי מובהקת.*

*קורולציית ספירמן (*Spearman*)*

*קורולציית ספירמן דומה לקורולציית פירסון (למעשה יש להן את אותה הנוסחה) – אבל במקום להשתמש בערך עבור כל דוגמא/מדד, היא משתמשת בדירוג שלו. לדוגמא:*

|  |  |  |  |
| --- | --- | --- | --- |
| ***תלמיד*** | ***מקצוע*** | ***ציון*** $(x\_{i})$ | ***דירוג*** $(R\left(x\_{i}\right))$ |
| ***אריק*** | *ביולוגיה של התא* | 100 | *1* |
| *אותות ומערכות ב'* | 50 | *3* |
| *פיזיקה 2ר'* | 70 | *2* |
| ***מיטל*** | *ביולוגיה של התא* | 30 | *3* |
| *אותות ומערכות ב'* | 90 | *2* |
| *פיזיקה 2ר'* | 100 | *1* |

$$ρ=\frac{1}{n}⋅\frac{\sum\_{i}^{n}\left(\left(R(x\_{i})-\overbar{R(x\_{i})}\right)⋅\left(R(y\_{i})-\overbar{R(y\_{i})}\right)\right)}{σ\left(R(x)\right)⋅σ\left(R(y)\right)}$$

*נחשב מרחקי ספירמן עבור השאלה מתחילת התרגול:*

spearman\_distance<-round(1-cor(t(grades),method="spearman"),2)



Heatmap(spearman\_distance, cluster\_columns = F, cluster\_rows = F, name = "Speaman\nDistance”, col = c("#fff9e3","#719d92","#003f5c"))



*כלומר, לפי קורולציית ספירמן הסטודנט שהכי קרוב לסטודנט 1 הוא סטודנט 4 – לו יש קורולציה מלאה ומרחק 0 (שימו לב הפונקציה מחזירה את המרחק ולא את הקורולציה).*

*אם נסתכל שוב על הגרף של הקורולציות נראה שהיחס בין הציונים של סטודנט 1 לסטודנט 4 הוא* ***לא לינארי*** *אבל הוא* ***כן מונוטוני****.*

*נסכם את מה שלמדנו על מרחקים:*

1. *מרחק אוקלידי נותן לנו מרחק לינארי בין דוגמאות/מדדים*
2. *קורולציית פירסון נותנת לנו קורולציה המתארת קשר לינארי בין דוגמאות/מדדים*
3. *קורולציית ספירמן נותנת לנו קורולציה המתארת קשר מונוטוני בין דוגמאות/מדדים*