**תרגול 9 – בחינת השערות**

בחלק הזה של הקורס נעסוק בניתוח של נתונים מרובי מימדים. בתרגול זה נדגים את חלק מהקשיים בהם אנו נתקלים בהם בעולם הרב-מימד.

נתחיל מלהסתכל על נתונים מרובי מימדים – איך הם נראים?

בקובץ ‘ritalin.mat’ ישנה מטריצה של ציוני בחינות. כל טור מייצג סטודנט (סה"כ 78 סטודנטים), וכל שורה מייצגת מבחן מסויים (40 בחינות סה"כ). חצי מהסטודנטים מטופלים בריטלין (עמודות 1-39) וחצי לא (עמודות 40-78). נרצה לדעת באילו בחינות יש הבדל בציוני הסטודנטים בין סטודנטים שלוקחים ריטלין לבין אלו שלא.

load("ritalin\_data.Rdata")

נציג את הדאטא בעזרת מפת חום (heatmap):

Graphical user interface, chart, application

Description automatically generated

סקלת הצבעים מצביעה על הציון. הגרף קצת קשה להבנה אבל אפשר לראות שיש מבחן אחד בו הציון הממוצע הוא מעל 100.

אבל כאן אנחנו לא יכולים להבין האם יש הבדלים בין קבוצות הסטודנטים ולכן נעמיק בניתוח הנתונים ונבצע מבחן t. להזכירכם, במבחן t אנו בודקים הם קיימים הבדלים בין שתי קבוצות. לשם כך אנו מגדירים שתי השערות מחקר (היפותזות):

**השערת האפס (H0 – null hypothesis) –** המניחה שממוצעי שתי הקבוצות שווים

**השערה אלטרנטיבית (H1 – alternative hypothesis) –** המניחה שממוצעי שתי הקבוצות לא שווים.

אלו הן השערות המחקר שלנו – מה שאנחנו בודקים לפני שיש לנו נתונים. אחרי שערכנו את הניסוי וקיבלנו נתונים, ייתכן והמציאות שונה ממה ששיערנו. ייתכנו ארבעה מצבים בין השערות המחקר למציאות:

|  |  |  |
| --- | --- | --- |
| **מציאות**  **החלטה** | **השערת האפס היא** | |
| **נכונה** | **לא נכונה** |
| **דחיית H0** | טעות מסוג 1 (false positive) | החלטה נכונה (true positive) |
| **קבלת H0** | החלטה נכונה (true negative) | טעות מסוג 2 (false negative) |

– ההסתברות לטעות טעות מסוג 1, – ההסתברות לטעות טעות מסוג 2

תוצאה חיובית היא לדחות את השערת האפס ותוצאה שלילית היא לא לדחות את השערת האפס. לפני שנבצע את המבחן נקבע שהוא רמת המובהקות הסטטיסטית – או ההסתברות לטעות טעות מסוג ראשון (לדחות את השערת האפס כאשר היא נכונה, false positive). אם נחזור לדוגמת ה-t-test, זה אומר שהסקנו כי ממוצעי הקבוצות שונים למרות שהם לא.

מבחן t לוקח את הממוצעים וסטיות התקן של הקבוצות, וקובע באמצעות נוסחה את ערך ה-p value של ההשוואה: הסיכוי לקבל את הנתונים שלנו אם במציאות היפותזת האפס נכונה. היפותזת האפס נדחית (והשערת המחקר מאוששת) אם מתקיים:

*ככל שרמת המובהקות שנדרוש () גבוהה יותר ( קטן – מובהקות גבוהה), כך הסיכוי לטעות מסוג ראשון יורד. מצד שני זה בדרך כלל יעלה את הסיכוי לטעות מסוג שני (הסיכוי שנפספס הבדלים אמיתיים). מקובל להשתמש ב- ומטה על מנת להגדיר רמת מובהקות גבוהה.*

*נמשיך בדוגמת הריטלין ונבצע מבחני t לכל אחת מהבחינות על מנת לראות במי מהן מבחינים בין שתי קבוצות הסטודנטים:*

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

}

#check for significant comparisons

sig\_tets<-which(data\_p < alpha)

#number of significant tests

num\_sig<-sum(data\_p < alpha)

נמצאו 11 מבחנים שבהם ההבדלים בין הקבוצות מובהקים, תחת אלפא של 0.05. אבל האם זה באמת אומר שיש סיכוי של 0.05 לטעות מסוג ראשון ב-11 המבחנים הללו?

השערות מרובות

כאשר אנו עושים השערות מרובות או בוחנים מספר היפותזות מאותם הנתונים, יש סיכוי לקבל תוצאות חיוביות שגויות (false positive או false discovery). הסיבה לכך היא שלפעמים מקבלים תוצאה חיובית במקרה, רק בגלל שערכנו מבחנים רבים.

למשל, אם ערכנו 200 מבחנים סטטיסטיים, השערת האפס נכונה עבור כל מבחן ויש לנו p value<0.05, נצפה ש-10 מתוך התוצאות האלו יהיו מובהקות רק במקרה:

כלומר, נקבל 10 תוצאות שהן "מובהקות" רק בגלל מקרה – הן בעצם false positive. אם כך, כאשר עושים מבחנים מרובים מגדילים את הסיכוי לטעויות מסוג 1. במילים אחרות - עבור כל מבחן בו דחינו את השערת האפס קיים סיכוי של עד 0.05 לטעות מסוג ראשון. ככל שכמות ההשוואות גדלה, כך הסיכוי שטעינו לפחות פעם אחת גדל, על אף ששמרנו על סיכוי נמוך לטעות בכל מבחן בפני עצמו.

**שאלה:** נניח ועשינו 40 מבחני t עם . מהי ההסתברות לעשות טעות מסוג 1?

משתמשים ב-Family-wise error rate (או FWER) על מנת לתאר את , רמת המובהקות עבור כל הניסוי. ה-FWER הוא ההסתברות לעשות גילוי שגוי אחד או יותר (= טעות מסוג 1):

***פתרון:*** *ההסתברות לעשות לפחות טעות אחת כאשר יש 40 מבחני t עם*  היא אם כך:

*כלומר, רוב הסיכויים שנעשה טעות מסוג 1!*

Chart

Description automatically generated*אם נסתכל על גרף הפונקציה של* FWER *נראה כי ככל שיש יותר מבחנים, הסיכוי לטעות טעות מסוג 1 הולך ושואף ל-1:*

*דרך להתמודד עם ההשוואות המרובות היא לתקן את ה-p-value על ידי תשלום של "קנס" לכל מבחן שנעשה. תיקון בונפרוני מנסה לשלוט ב-FWER אך הוא תיקון מחמיר ביותר – רמת המובהקות של הניסוי מותאמת בו לכמות ההשוואות שעשינו. עבור כל השוואה, הקבוצות יוכרזו כשונות רק אם ה-p value יקיים:*

כאשר m הוא כמות ההשוואות.

למשל, אם ביצענו 100 מבחנים סטטיסטיים עם , נקבל שכל אחד מהמבחנים צריך  
 על מנת להיות מובהק.

אם כן – ביצענו בהצלחה מרובה את הורדת הטעויות מסוג 1, אבל מצד שני העלנו משמעותית את הסיכוי לטעות מסוג 2 (פספוס של הבדלים אמיתיים, false negative).

נחזור לדוגמא של הריטלין – יש לנו 40 מבחנים ולכן , נבדוק אילו מהמבחנים נשאר מובהק:

alpha<-0.05/40

*קיבלנו שאחרי תיקון בונפרוני, רק 4 מבחנים הבחינו באופן מובהק בין קבוצות הסטודנטים.*

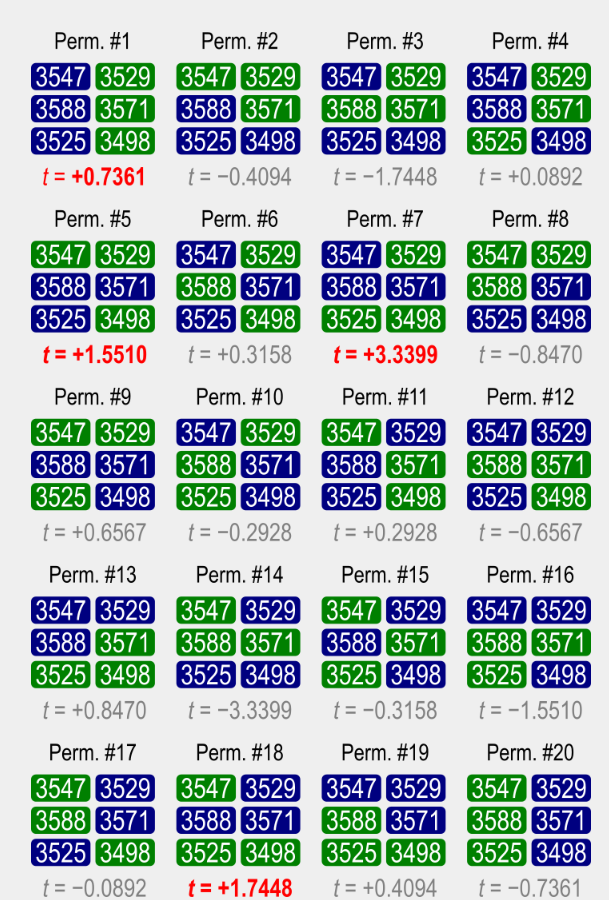
*פרמוטציות*

בעיה נוספת איתה מתמודדים כאשר מנתחים כמות רבה של נתונים היא שצריכים להניח הרבה מאוד הנחות כדי להשתמש במבחנים הסטטיסטיים. כדי להתמודד עם הבעיה הזו ניתן לעשות *פרמוטציות.*

*נניח ויש לנו קבוצה של אנשים עם סוכרת וקבוצה לא, ואנו רואים הבדל של ברמות הכולסטרול הממוצעות בין שתי הקבוצות. האם זהו הבדל משמעותי? כדי לבדוק זאת אפשר לערבל באופן אקראי את ה"תוויות" עבור כל אחד מהאנשים, לחלק לשתי קבוצות בגודלן המקורי ולבדוק שוב האם יש הבדל (מבחן t, p-value...) בממוצע הכולסטרול בין קבוצת ה"סכרתיים" לקבוצת ה"לא סכרתיים".*

*בסוף התהליך, נבדוק כמה פעמים מתוך הפעמים שבדקנו קיבלנו סטטיסטי גדול מהסטטיסטי של המבחן האמיתי. ה-p-value יינתן על ידי חלוקת מספר הפעמים הנ"ל במספר הפרמוטציות:*

*בדוגמא הבאה מוצג משקל ההיפוקמפוס אצל 6 אנשים. 3 עם מחלת אלצהיימר (ירוק) ו-3 בריאים (כחול). ערכנו מבחן t וקיבלנו הבדל בין הקבוצות עם . על מנת לבדוק את ההשערה, ערכנו 20 פרמוטציות שתוצאותיהן מוצגות בתמונה. חשבו את ה- p-value עבור המבחן האמיתי.*



***פתרון:*** *ניתן לראות כי ב-3 מתוך 20 הפרמוטציות קיבלנו t גדול או שווה ל-t של המבחן האמיתי. כלומר, ה-p-value שלנו הוא:*

*כלומר, אין הבדל משמעותי בין הקבוצות.*