ב"ה

המידות המסכמות לקִרבה

מאת דורון ויצטום

"מידת הקִרבה המכוילת" מודדת את הסיכוי למפגש עבור זוג ביטויים בודד. באמצעותה מקבלים קבוצה של מספרים עבור קבוצה של זוגות. אך כיצד נעריך מה הסיכוי לקבל קבוצה "כזאת" של מספרים?

אנו זקוקים אפוא למספר מסכם, אשר ייתן את ההסתברות למדגם כולו. בעצם, זו מידה ל"נטייה הכוללת לקרבה" עבור כל הזוגות במדגם. להלן נגדיר שתי מידות (סטטיסטיקות) כאלו, שהשתמשנו בהן במחקרנו.

א. הגדרת מידת "הנטייה הכוללת לקִרבה" P_1.

לפי מידה זו מונים את מספר התוצאות ב"אזור ההצלחה", אשר הוגדר (שרירותית) כמרווח בין 0 ל- 0.2, ומחשבים מה הסיכוי לקבל באקראי את הערך המתקבל.

המדגם העומד לבדיקה הוא קבוצה של זוגות ביטויים. "מידת הקרבה המכויילת" של כל זוג ביטויים (w,w') ניתנת לחישוב על ידי $c(w,w')$. כך מקבלים N מספרים, שכל אחד מהם הוא בין 0 ל- 1. נניח שמספר הזוגות (w,w') עבורם $c(w,w')\le 1/5$ הוא k. נגדיר

$$P_1 \equiv \sum_{j=k}^{N} \binom{N}{j} \left(\frac{1}{5}\right)^j \left(\frac{4}{5}\right)^{N-j} \qquad (5.1)$$

כדי להבין הגדרה זאת, נשים לב לכך, ש**אם** המספרים $c(w,w')$ הם משתנים אקראיים בלתי-תלויים המתפלגים בצורה אחידה (אוניפורמית) בין 0 ל- 1, אזי P_1 היא ההסתברות, כי לפחות k מ-N המספרים - קטנים או שווים ל- 0.2.

בהערכת תוצאות הניסוי הגדול הראשון (והשני) אכן הנחנו אחידות (אוניפורמיות) ואי תלות. התברר, שהנחה זו לא היתה מוצדקת. בהמשך המחקר השתמשנו במידה זו, אך לא הנחנו

ואף לא עשינו כל שימוש בהנחה של אחידות ואי-תלות. לכן, למרות ש-P_1 מכוילת כהסתברות, היא משמשת רק כמדד סידורי. P_1 מודדת את מספר זוגות הביטויים במדגם, שבהם בני הזוג "קרובים למדי" זה לזה (כלומר, $c(w,w') \le 1/5$), ועם זאת גם מביאה בחשבון את גודל המדגם כולו. מידה זו מאפשרת לנו להשוות את "הנטייה הכוללת לקרבה" במדגמים שונים; בייחוד במדגמים הנוצרים על ידי הפרמוטציות במבחן הרנדומיזציה (ראה "צופן בראשית", פרק י"ג ונספח 9).

נשים לב, כי הסטטיסטיקה P_1 מתעלמת מכל ערכי $c(w,w')$ הגדולים מ- 0.2, ומעניקה אותו המשקל לכל ערכי $c(w,w')$ הקטנים מ- 0.2. כלומר, אנו מתמקדים במפגשים המצליחים ללא הבחנה באיכותם, ואיננו מתעניינים לדעת באיזו מידה נכשלו אלה שלא הצליחו.

ב. הגדרת מידת "הנטייה הכוללת לקרבה" P_2.

המידה P_2 נבנתה כך, שהיא רגישה לגודלם של **כל** המספרים $c(w,w')$. המשמעות של P_2 היא, ש**אם** המספרים $c(w,w')$ הם משתנים אקראיים בלתי-תלויים המתפלגים בצורה אחידה בין 0 ל- 1, אזי P_2 היא ההסתברות, שמכפלת ערכי $c(w,w')$ תהיה קטנה כפי שהיא, או קטנה מזה.

הגדרת מידה זו מסובכת יותר. ראשית, אנו מחשבים את המכפלה $\prod c(w,w')$, שבה נכפלים N המספרים $c(w,w')$ שחושבו עבור הזוגות המדגם. אחר כך, אנו מגדירים

$$P_2 \equiv F^N\left(\prod c(w,w')\right) \qquad (5.2)$$

כאשר
$$F^N(X) \equiv X\left(1 - \ln X + \frac{(-\ln X)^2}{2!} + \cdots + \frac{(-\ln X)^{N-1}}{(N-1)!}\right)$$

כדי להבין הגדרה זו, נשים לב כי אם $x_1, x_2, \ldots, x_N$ הם משתנים אקראיים בלתי-תלויים המתפלגים בצורה אחידה בין 0 ל- 1, אזי ההתפלגות של מכפלתם $X \equiv x_1 x_2 \cdots x_N$ ניתנת על ידי

$$\Pr(X \le X_0) = F^N(X_0)$$

[הדבר נובע מתוצאה (3.5) של פלר[1] מכיוון ש- $-\ln x_i$ מתפלגים באופן אקספוננציאלי, וגם $-\ln X = \sum_i \left(-\ln x_i\right)$.]

גם לגבי מידה זו, השתמשנו בהנחה של אחידות ואי תלות בהערכת תוצאות הניסוי הגדול הראשון (והשני). אך התברר, שהנחה זו לא היתה מוצדקת. בהמשך המחקר השתמשנו במידה זו, אך לא הנחנו ואף לא עשינו כל שימוש בהנחה של אחידות ואי-תלות. לכן, למרות ש-P_2 מכויל כהסתברות, הוא משמש רק כמדד סידורי, המאפשר לנו להשוות את "הנטייה הכוללת לקרבה" במדגמים השונים.

ג. הגדרת מידות "הנטייה הכוללת לקִרבה" P_3 ו- P_4.

עבור המדגם השני, שבשבילו תוכנן מבחן הראנדומיזציה (ראה "צופן בראשית" פרק י"ג), הוגדרו שתי סטטיסטיקות נוספות: P_3 ו- P_4. למעשה, היו אלו בדיוק P_1 ו- P_2 בהתאמה, שהוגדרו עבור מדגם חלקי, אשר יוגדר להלן.

הצורך במדגם החלקי יובן, אם נתבונן ברשימת השמות והכינויים של המדגם השני (ראה קובץ "המדגם השני" טבלה 2), בכינויים המופיעים בטור "רבי ...": הכינויים יכולים להיות משותפים לכמה אישים. ואכן, הרשימה כללה ארבעה "רבי אברהם", שלושה "רבי דוד", ארבעה "רבי חיים" וכן הלאה. ברור, שגם *לאחר* שיבוש המדגם על ידי צימוד אקראי המצמיד תאריכים לאישים, יוותרו זוגות "כינוי – תאריך" מן המדגם המקורי (למשל, "רבי דוד" אחד "יקבל" את תאריכיו של "רבי דוד" אחר). כך, שחלק מן המדגם המשובש כלל לא יהיה משובש!

משום כך, הגדרנו מדגם חלקי שאינו כולל את הכינויים "רבי ..." שבטבלה הנ"ל. אפשר להגדיר מדגם חלקי כזה גם עבור המדגם הראשון, וכן לגבי כל מדגם בעל נתונים מסוג זה.

המידה P_1 המיושמת לגבי המדגם החלקי תיקרא P_3, ואילו המידה P_2 המיושמת לגבי המדגם החלקי תיקרא P_4.

[1] Feller, W. (1966). *An Introduction to Probability Theory and Its Applications* 2. Wiley, New York.