נספח א3

המידות המסכמות לקִרבה

"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 לקִרבה" עבור כל הזוגות במדגם. להלן נגדיר שתי מידות (סטטיסטים) 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

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

כדי להבין הגדרה זאת, נשים לב לכך, **שאם** 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 מבחני רנדומיזציה.

[הערה: החישוב ב(1), לפי ההתפלגות הבינומלית, מסובך למדי. בהעדר תוכנה מתאימה, מעדיפים להשתמש בקירוב הנורמלי להתפלגות הבינומלית. כך עשינו בדיווחים[1] הראשונים שלנו על תוצאות שני הניסויים הגדולים שיתוארו בהמשך. ספרנו כמה תוצאות שערכן קטן או שווה ל-p=0.2. סימננו מספר זה ב- S. ספרנו כמה סטיות תקן יש בין S לבין הממוצע הצפוי באקראי (Np). מידה זו נסמן ב- P'_1. לאחר מכן, כאשר הוכנה עבורנו תוכנה המחשבת ישירות את (1), נקבנו בערך המדויק המתקבל בחישוב ההתפלגות הבינומלית.]

נשים לב, כי המידה 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 לקִרבה" P_2

המידה P_2 נבנתה כך, שהיא רגישה לגודלם של **כל** המספרים $c(w,w')$. המשמעות של P_2 היא, **שאם** 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

$$P_2 \equiv F^N\left(\prod c(w,w')\right) \qquad (2)$$

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[2] מכיוון ש- $-\ln x_i$ מתפלגים באופן אקספוננציאלי, וגם $-\ln X = \sum_i(-\ln x_i)$.]

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

---

[1] D. Witztum, E. Rips, and Y. Rosenberg, *Equidistant Letter Sequences in the Book of Genesis*. Preprint. 1986.
D. Witztum, E. Rips, and Y. Rosenberg, *Equidistant Letter Sequences in the Book of Genesis*. Preprint. 1988.

[2] W. Feller, *An Introduction to Probability Theory and Its Applications* 2. Wiley, N.Y. 1966.

מכויל כהסתברות, הוא משמש רק כמדד סידורי, המאפשר לנו להשוות את "הנטייה הכוללת לקרבה" במדגמים השונים.

ג. השוואה בין P_1 ו- P_2

1. כל אחת משתי המידות P_i משמשת מעין "גלאי" להצפנה המשוערת.

- P_1: לפי מידה זו מונים את מספר התוצאות $c(w,w')$ ב"אזור ההצלחה", אשר הוגדר א-פריורי כמרווח בין 0 ל- 0.2, ומחשבים מה הסיכוי לקבל באקראי את הערך המתקבל.
- P_2: מידה זו נבנתה כך, שהיא רגישה לגודלם של **כל** המספרים $c(w,w')$.

2. לכל סטטיסטי, P_1 או P_2, יש יתרונות וחסרונות.

(א) למשל, ל- P_1 יש יתרון, שאין השפעה של "גודל הכשלון" של הזוגות "הנכשלים". אנחנו מחפשים הצפנה – זוגות "מוצפנים". לא מעניין אותנו כיצד בדיוק מתנהגים אותם זוגות שאינם מוצפנים. עניין זה עצמו הוא חסרון של P_2, הרגישה לתוצאות באזור הכשלון (סמוך ל-1).

(ב) לעומת זאת ל- P_2 יש יתרון, שהיא רגישה לחדות ההצלחה. עניין זה עצמו חסר ל- P_1, שעבורה כל תוצאה ב"אזור ההצלחה" – ערכה שווה.

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

עבור המדגם השני, שבשבילו תוכנן מבחן הפרמוטציות (ראו פרק יט), הוגדרו שתי מידות 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 (ראו נספח א6 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.

נסכם כל זאת בצורה מדויקת:

1. המדגם השני, בנוי מזוגות ביטויים (w, w'). בכל זוג, w הוא שם אישיות (או כינויה) מקבוצת אישים (גדולי תורה) ו- w' - תאריך הלידה או הפטירה שלה.
2. השמות והכינויים של כל אישיות הם משני סוגים: שמות וכינויים המיוחדים לאותה אישיות, והכינוי הסטנדרטי רבי "פלוני" (כאשר "פלוני" הוא שמו העברי הפרטי), המשותף לכמה אישים.

לדוגמא: לאישיות #1 ברשימה, כינוי "אישי" הראב"י, וכינוי סטנדרטי רבי אברהם. הכינוי רבי אברהם ניתן לכל חכם ששמו הפרטי אברהם, ולכן הוא משותף לארבעת האישים הראשונים בקבוצת האישים[3].

3. לכן, ניתן להציג את המדגם השני השלם, *LIST 2*, כאיחוד של שני תת-מדגמים:

- תת-המדגם *L2* הבנוי מאותם זוגות ביטויים (w, w'), שבהם w הוא מן השמות והכינויים המיוחדים.
- תת-המדגם *L'2* הבנוי מאותם זוגות ביטויים (w, w'), שבהם w הוא הכינוי רבי פלוני.

$$LIST\ 2 = \text{המדגם השני השלם} = L2 + L'2$$

אבחנה זו נעשתה לפני ביצוע מבחן הפרמוטציות על ידינו, כפי שנתבאר במאמרנו[4] (עמ' 436 בסופו – 437).

4. מבחן הפרמוטציות, שפרטיו סוכמו מראש בין ישראל אומן לפרסי דיאקוניס, נועד למדוד את מובהקותם של ארבעה סטטיסטים: שני סטטיסטים לגבי הרשימה כולה *LIST 2* ואותם שני סטטיסטים לגבי *L2*.

במלים אחרות: ישנם שני אופראטורים, $\boldsymbol{P_1}$ ו- $\boldsymbol{P_2}$.

היישום שלהם על המדגם השלם נותן את P_1 ו- P_2: $P_1 \equiv \boldsymbol{P_1}(LIST\ 2),\ P_2 \equiv \boldsymbol{P_2}(LIST\ 2)$

והיישום שלהם על תת-המדגם *L2* נותן את P_3 ו- P_4: $P_3 \equiv \boldsymbol{P_1}(L2),\quad P_4 \equiv \boldsymbol{P_2}(L2)$

---

[3] כינויים סטנדרטים אלה נמצאים בעמודה מיוחדת בטבלת השמות והכינויים שהכין המומחה החיצוני, פרופסור שלמה זלמן הבלין ראו בנספח א6 טבלה 2.

[4] המאמר:

D. Witztum, E. Rips, and Y. Rosenberg, *Equidistant letter sequences in the Book of Genesis*. Statist. Sci. 9 No. 3 (1994), pp. 429-438.

נמצא גם באתר "צופן בראשית" בקישור: http://www.torahcode.co.il/pdf_files/pub/wrr.pdf