נספח א5

טעות פשוטה

בנספח א2 ("מדידת מפגשים") הגדרנו את "מידת הקִרבה". זו מקבלת ערך גדול יותר אם המד"שים הנפגשים מינימליים יותר, קרובים יותר זה לזה ו"לא מפוזרים" במידה רבה יותר. שם גם הגדרנו את "מידת הקִרבה המכוילת" המודדת מה הסיכוי, של"מידת הקִרבה" ערך כה גדול. כפי שהוסבר שם באריכות, "מידת הקרבה המכוילת" היא הדרוג ב"מרוץ" בין המד"שים למדכ"שים (מלים בדילוגים כמעט שווים). ככל שערכה נמוך יותר, הסיכוי שהמפגשים אכן נוצרו במקרה – קטן יותר.

להפתעתנו, הסטטיסטיקאי הנודע, פרופסור פרסי דיאקוניס כתב במכתבו הראשון (ראו בפרק טז), כי לא הצליח להבין מדוע נזקקנו להגדיר את "מידת הקִרבה המכוילת". דיאקוניס חשב בטעות, כי ניתן להשוות ישירות את "מידת הקִרבה" של זוג ביטויים אחד ל"מידת הקִרבה" של זוג ביטויים אחר. מכאן הסיק, כי אפשר ליישם שיטות סטטיסטיות סטנדרטיות המבוססות על השוואה, ישירות על קבוצת ערכי "מידת הקִרבה" שנתקבלה בניסוי הגדול. לכן, הוא הציע באותו מכתב מבחן מסוים המבוסס על השוואה ישירה של "מידת הקרבה" של זוגות ביטויים. כפי שנראה להלן, זו טעות חמורה. למרבה האירוניה, העובדה שאנחנו לא עשינו כך – עוררה בו חשד, ששיטת ההשוואה לדילוגים כמעט שווים, "נתפרה" כדי לייצר מובהקות שאינה אמיתית!...

מבקר אחר של מחקרנו, פרופסור אברהם הסופר (אף הוא סטטיסטיקאי), חזר על אותה שגיאה כעשר שנים מאוחר יותר, באופן בלתי תלוי, ואף הוסיף נופך משלו. הוא הצביע, כי לכאורה הגדרת "מידת הקרבה המכוילת" מניבה מוזרויות ביחס להגדרת "מידת הקִרבה". לדוגמא הוא מביא "פרדוקס":

"מידת הקִרבה המכוילת" של זוג הביטויים "הגאון" – "ט"ו ניסן" שווה ל- 0.076. פירושו של דבר, כי קיים סיכוי קטן למדי לקבל במקרה מפגש כל כך מוצלח. לעומת זאת, "מידת הקִרבה המכוילת" של זוג הביטויים "רבי משה" – "בי"ח איר" היא 0.4. כלומר, היא גרועה פי 5.3 מזו של הזוג הראשון, ומצביעה שהמפגש של הזוג הראשון טוב מזה של השני. אבל, אם בוחנים את "מידת הקִרבה" מקבלים תמונה הפוכה: "מידת הקִרבה" של הזוג השני, היא טובה פי 4.5 מזו של הזוג הראשון – ומצביעה שהמפגש של הזוג השני טוב מזה של הראשון!

ננסה להבהיר את שגיאתו של דיאקוניס ואת טעות הסופר באמצעות משל.

היוונים הקדמונים רצו להוכיח את עליונותם בריצה. לטענתם, עליונות הרץ היווני היא תכונה מוּלדת, ולכן היא קיימת בכל הגילים. הם יזמו סדרה של 99 תחרויות ריצה.

- בכל מרוץ היו אמורים להשתתף 125 רצים: אחד מהם יווני ושאר 124 הרצים הם בני עמים ובני שבטים אחרים.

- המרוצים התנהלו לפי שנתונים: במרוץ הראשון השתתפו רצים בני שנתיים, במרוץ השני – בני שלוש שנים, בשלישי – בני ארבע, וכן הלאה, עד כי במרוץ ה-99 השתתפו רצים בני 100 שנה.
- לכל אחד מן הרצים המשתתפים נרשמה "מידת המהירות" באמצעות שעון חול משוכלל.
- בכל מרוץ דרגו את הרצים לפי סדר הגעתם למטרה. ואז, הגדירו את "מידת המהירות המכוילת" של המשתתף היווני כך: הדרוג שלו, מחולק במספר המשתתפים בפועל במרוץ. למשל, אם היווני הגיע למקום השני מתוך 109 משתתפים בפועל (16 רצים לא התייצבו לתחרות או מתו בדרך), אזי "מידת המהירות המכוילת" שלו היא 2/109.
- תוצאות האולימפידה הזאת היו, אם כן, 99 מספרים הקטנים או שווים ל-1. מארגני האולימפידה צריכים היו לקבוע, אם התוצאות אכן מאשרות את עליונות היוונים כטענתם – ובאיזו רמת מובהקות סטטיסטית.
- ניתוח התוצאות יכול להיעשות על ידי "המידות הכוללות לקִרבה", שהוגדרו בנספח א3 ("קִרבה כוללת"). אך לא נעסוק בכך עכשיו, רק נדווח כי באותו מעמד אכן הוכחה טענת היוונים במובהקות ניכרת, ובטקס נעילת האולימפידה הוכרז על כך בחגיגיות.

והנה, קם הנציג המצרי – בכיר הנציגים הזרים – וערער על התוצאות ועל המסקנות. לטענתו היה פגם בסיסי בהגדרת "מידת המהירות המכוילת" – פגם הגורם לפרדוקסים משונים. לדוגמא, המצרי הצביע על כך, שבתחרות לבני 33 שנים, הגיע הרץ היווני רק למקום ה-50 מתוך 125 רצים, כך ש"מידת המהירות המכוילת" שלו היא 50/125. לעומת זאת, בתחרות לבני 99 הגיע הרץ היווני למקום הראשון מתוך 101 רצים, כך ש"מידת המהירות המכוילת" שלו היא 1/101, כלומר, הישג שהוא טוב יותר פי 40 מזה של הרץ בן ה-33.

"הרי זה אבסורד!" – צעק המצרי בהתרגשות – "כל אחד יודע שהמצב הפוך: 'מידת המהירות' של הרץ בן ה-33 טובה פי 50 מזו של בן ה-99!"

מארגני האולימפידה הסבירו למצרי הנרגש, כי טעות בידו. לכל מרוץ דרגת קושי ייחודית, התלויה בכושר הגופני של המתחרים. יש משמעות ל"מידת המהירות" של הרצים ביחס לרצים "דומים": היא נועדה לבדוק, אם לרץ היווני יכולת ריצה טובה במיוחד בהשוואה לרצים בני אותו שנתון, שלהם תכונות גופניות דומות. היא קובעת את הדרוג, הדרוש להגדרת "מידת המהירות המכוילת", שהיא המדד ל"יכולת המיוחדת". לעומת זאת, אין טעם להשוואה בין "מידת המהירות" של רץ בן 33, לזו של רץ בן 99, או לזו של רץ בן שנתיים: אי אפשר ללמוד מהשוואה כזו על יכולת הריצה המיוחדת של הרץ בן 33, אלא רק על השפעות הזקנה המופלגת על בן ה-99, או על ההשפעה של רגלים קצרצרות ומוטוריקה שלא בשלה בבן השנתיים.

איננו יודעים אם הנציג המצרי השתכנע או לא. נניח לו ונחזור לנמשל.

הנמשל הוא - 152 "המרוצים" שנערכו במסגרת הניסוי הגדול בין המד"שים לשאר המדכ"שים. גם כאן, לכל "מרוץ" היתה דרגת קושי ייחודית. כאן לא גיל המתחרים קבע את דרגת הקושי, אלא זוג הביטויים הנמדדים קבע זאת, כפי שנסביר מיד.

למשל, אם "מלה א" בזוג מלים היא בת 5 אותיות שכיחות, צפוי כי המד"שים שלה יופיעו בדילוג קטן מ-10. אם "מלה ב" בזוג היא מלה בת 7 אותיות, צפוי כי המד"שים שלה יופיעו בדילוג

של כמה אלפים. במקרה כזה, קל יחסית למד"ש מינימלי בעל דילוג גדול (השייך ל"מלה ב") "ללכוד" מד"ש מינימלי בדילוג קטן (השייך ל"מלה א") במפגש קרוב. המד"ש בעל הדילוג הגדול קובע את הטבלה הדו-ממדית, ולכן יחסית הוא אינו מפוזר. המד"ש בעל הדילוג הקטן אף הוא אינו מפוזר (כי הדילוג קטן ולאורך השורה בטבלה). כך מקבלים בקלות יחסית מפגש קרוב ולא מפוזר.

נדגים זאת: הטבלה הבאה, ובה 1700=[6801/4] טורים, נקבעה על ידי המד"ש המינימלי של המלה "הסתברות" בכל ספר בראשית:

טבלה 1

****
ברואתויצחקוישמעאלבניואלמערתהמכ
אמראכהנהאשתכהואואיכאמרתאחתיהוא
בקלילאשראנימצוהאתכלכנאאלהצאנוק
ובלבויקרבוימיאבלאביואהרגהאתיעק
למקמהויאמרלהמיעקבאחימאינאתמויא
ליהיעקבותהרבלההותלדליעקבבנותאמ
שבימהפרידיעקבויתנפניהצאנאלעקדו
ספתהלביתאביכלמהגנבתאתאלהיויעני
מהלכלקראתכוארבעמאותאישעמוויירא
שפחותהנהויולדיהנותשתחויניותגשגמל
אתבנתמנקחלנולנשימואתבנתינונתנל
ההפילגשאבויושמעישראלויהיובניי
אדומבלעבנבעוראושמעירודנהבהויתמב
תואתכתנתהפסימאשרעליויקחהויושל
אלאתאנשימקמהלאמראיההקדשההואבעי
נחנובעינישרביתהסהרויתנשרביתהסה
יהיבבקרותפעמרוחוויושלחויקראאתכל
צמצרימויסרפרעהאתטבעתומעלידוויית
משמרשלשתימימויאמראלהמיוספביומה
קשנואמלאהביאתיואליכוהצגתיולפני
הרחיקוויוספאמרלאשרעלביתוקומרדפ
יולאיכלואחיולענותאתוכינבהלומפנ
ימהיעקבוכלזרעואתובניוובניבניוא
עמדהולפניפרעהויברכיעקבאתפרעהוי
לוישבעלהמטהויאמריעקבאליוספאלשד
האמוי, הסתברות
****

על פני הטבלה מופיע מפגש קרוב ולא מפוזר עם מד"ש מינימלי של המלה המלאכותית "האמוי", בת 5 אותיות שכיחות (להלן יוסבר איך הגענו אליה). למרות שהמפגש נראה "פוטוגני" – ואכן "מידת הקִרבה" שלו גדולה – הוא אינו מובהק. בתחרות בין המד"שים למדכ"שים מתברר, כי הדרוג של מפגש זה הוא 19 מתוך 55 מתחרים, ולכן "מידת הקרבה המכוילת" של מפגש נאה זה (ראו בסוף נספח א2) היא רק 19/55. הסיבה לכך, כי גם למתחרים – המדכ"שים – אותה דרגת קושי (במקרה זה – אותה קלות) ליצור מפגש כזה או טוב ממנו. מקרה כזה מקביל למרוץ של בני 33 במשל הנ"ל, ואולי אפילו למרוץ של בני 25.

לעומתו, נציג עתה "מרוץ" בדרגת קושי המקבילה (אולי) למרוץ של בני 70. הדבר יקרה כאשר שני הביטויים הם בני 7 אותיות.

נדגים זאת: הטבלה הבאה, ובה 680=[6801/10] טורים, נקבעה על ידי המד"ש המינימלי של המלה "הסתברות" בכל ספר בראשית:

אמיוהלה, הסתברות

על פני הטבלה מופיע מפגש קרוב פחות ומפוזר יותר מן הקודם, עם מד"ש מינימלי של המלה המלאכותית "אמיוהלר", שהיא בת 7 אותיות שכיחות (להלן יוסבר איך הגענו אליה). למרות שהמפגש **אינו** נראה "פוטוגני" – ואומנם "מידת הקִרבה" שלו קטנה פי 4 מזו של המפגש הקודם – הוא מובהק בהרבה מקודמו. כאשר עורכים תחרות בין המד"שים למדכ"שים מתברר, כי הדרוג של מפגש זה הוא 7 מתוך 55 מתחרים, ולכן "מידת הקִרבה המכוילת" של המפגש היא 7/55. הסיבה לכך היא, כי קשה ליצור מפגש כזה כאשר המד"שים הנפגשים הם כה "מגושמים" (הם מייצגים ביטויים בני 7 אותיות). במצג זה הפגינו אפוא המד"שים "יכולת מיוחדת" ביחס למתחרים (המדכ"שים): דבר זה נחשב להישג טוב עבור מד"שים כל כך מגושמים, בדומה למשל, שלבן 70 ייחשב הישג טוב כאשר ירוץ במהירות הגרועה רק פי 4 מזו של מתחרה בינוני במרוץ של בני ה-33.

בדומה, מפגש המקביל לרץ המצטיין במרוץ של בני ה-99, יכול להתקבל במפגש של המלה "הסתברות" עם ביטוי בן 8 אותיות...

--- --- ---

עד כאן טיפלנו רק בדרגת הקושי של האפשרות למפגשים קרובים ולא מפוזרים בין מד"שים, המושפעת מ"מגושמות" המד"שים. כדאי לציין, כי קיים גורם חשוב נוסף, המשנה את דרגת הקושי במפגשים בין ביטויים ארוכים: הכוונה למספר המד"שים המייצגים את הביטויים. לדוגמא, ביטוי בן 8 אותיות צפוי להופיע בספר בראשית מספר פעמים מועט בדילוג שווה של אותיות, בדרך כלל רק פעם אחת בלבד (במקום כ-10 מד"שים המייצגים בדרך כלל את הביטוי בסכמת המדידה שלנו – ראו נספח א2). ולכן, במרוץ הנערך עבור זוג ביטויים, שאחד מהם בן 8 אותיות, ישתתפו מד"שים מועטים. נובע מכך, כי "מידת הקִרבה" – שהיא סכום התרומות של מפגשי המד"שים (ראו נספח א2) – תקטן באופן משמעותי.

אם כן טעותם הבסיסית של דיאקוניס והסופר היתה, שלא הבחינו כי "מרוצים" הנקבעים על ידי זוגות ביטויים שונים, אינם עומדים כלל באותה דרגת קושי, ולכן "מידת הקִרבה" במרוץ אחד אינה בת-השוואה לזו של מרוץ אחר. לפי טעותם, ערך התוחלת של "מידת הקִרבה" אינו תלוי בזוג הביטויים הנמדד. הרי זה כאילו טענו שערך התוחלת של "מידת המהירות" במרוץ מסוים באולימפידה (במשל שלנו), אינו תלוי בגיל הרצים!...

לפי דברינו עד כה, ניתן לחזות מראש, כי ערך "מידת הקִרבה" במפגשים עם מלה כמו "הסתברות" יהיה גדול יחסית כשהמלה השניה קצרה, אך קטן יחסית – כשהמלה השניה ארוכה. אפשר להעמיד זאת במבחן הניסוי.

ננסה להפגיש ביטויים משלושה סוגים - A, B ו- C - עם המלה "הסתברות".

סוג A: ביטויים בני 5 אותיות השכיחות ביותר בספר בראשית. כך מובטח כי צפוי שהם יופיעו בדילוג קצר. כדי לקבל קבוצה גדולה כזאת יצרנו את הביטוי "אהוימ", המורכב מחמש האותיות השכיחות ביותר בספר בראשית, שסודרו כאן לפי מיקומן בסדר האלפבית. ניתן לסדר אותיות אלה ב-120 אופנים שונים. כיוון שאין אנו מבדילים בין דילוג קדימה לדילוג אחורה, הרי לעניין הדילוגים הביטוי "אהוימ" והביטוי "מיוהא" מהווים אותו הביטוי. לכן, סך כל הביטויים השונים הוא 120/2=60.

סוג B: לכל אחד מן הביטויים מסוג A, נוסיף "לר" בסופו. כך שבמקום "אהוימ" – יתקבל "אהוימלר". האותיות "לר" נבחרו משום ש"ל" היא האות הששית מבחינת השכיחות בספר בראשית, ו"ר" – השביעית. כך קבלנו 60 ביטויים שונים בני 7 אותיות.

סוג C: לכל אחד מן הביטויים מסוג B, נוסיף "ב" בסופו, במקום "אהוימלר" – "אהוימלרב". האות "ב" נבחרה משום שהיא האות השמינית מבחינת השכיחות בספר בראשית. כך קבלנו 60 ביטויים שונים בני 8 אותיות.

שימו לב, כי בנית קבוצות הביטויים מאותן האותיות מבטיח, כי דרגת הקושי דומה היא עבור כל הביטויים השייכים לאותו הסוג.

בטבלה הבאה ציינתי בצד כל אחד מן הביטויים את ערכה של "מידת הקִרבה" במפגש עם המלה "הסתברות". לנוחות ההצגה הוכפל כל מספר פי 100. בטור האחרון סימנתי ב"-" את המקרים שבהם הביטוי אינו מופיע בדילוג שווה.

טבלה 3

	A	Ω×100	B	Ω×100	C	Ω×100
1	אהוימ	1.403	אהוימלר	1.118	אהוימלרב	0.325
2	המאוי	1.810	המאוילר	1.060	המאוילרב	0.453
3	המאיו	2.483	המאיולר	1.556	המאיולרב	-
4	המואי	6.842	המואילר	0.726	המואילרב	-
5	המיאו	3.517	המיאולר	1.001	המיאולרב	-
6	האמוי	3.100	האמוילר	0.663	האמוילרב	0.231
7	האמיו	5.124	האמיולר	1.600	האמיולרב	-
8	האומי	6.678	האומילר	0.842	האומילרב	0.490
9	האוימ	9.158	האוימלר	1.023	האוימלרב	0.626
10	האימו	4.038	האימולר	1.144	האימולרב	0.161
11	האיומ	1.379	האיומלר	0.989	האיומלרב	0.447
12	הומאי	1.732	הומאילר	0.702	הומאילרב	0.588
13	הואמי	3.998	הואמילר	0.731	הואמילרב	0.400
14	הואימ	0.601	הואימלר	1.635	הואימלרב	0.423
15	הויאמ	3.506	הויאמלר	1.035	הויאמלרב	0.386
16	הימאו	4.193	הימאולר	1.077	הימאולרב	0.255
17	היאמו	0.548	היאמולר	1.023	היאמולרב	0.295
18	היאומ	3.260	היאומלר	1.215	היאומלרב	0.217
19	היואמ	1.987	היואמלר	1.006	היואמלרב	-
20	אמהוי	1.608	אמהוילר	0.850	אמהוילרב	0.351
21	אמהיו	2.353	אמהיולר	1.303	אמהיולרב	0.562
22	אמוהי	1.488	אמוהילר	1.461	אמוהילרב	-
23	אמויה	2.934	אמויהלר	0.833	אמויהלרב	0.402
24	אמיהו	10.600	אמיהולר	1.392	אמיהולרב	0.399
25	אמיוה	1.605	אמיוהלר	1.088	אמיוהלרב	0.106
26	אהמוי	1.898	אהמוילר	0.701	אהמוילרב	0.192
27	אהמיו	10.518	אהמיולר	1.029	אהמיולרב	0.632
28	אהומי	2.775	אהומילר	1.019	אהומילרב	-
29	אהימו	1.403	אהימולר	1.015	אהימולרב	0.233
30	אהיומ	4.816	אהיומלר	0.969	אהיומלרב	1.304
31	אומהי	24.714	אומהילר	0.541	אומהילרב	0.469
32	אומיה	1.523	אומיהלר	1.164	אומיהלרב	0.399
33	אוהמי	2.781	אוהמילר	1.215	אוהמילרב	0.318
34	אוהימ	1.877	אוהימלר	1.625	אוהימלרב	0.750
35	אוימה	1.071	אוימהלר	0.580	אוימהלרב	0.360
36	אויהמ	0.906	אויהמלר	1.091	אויהמלרב	0.381
37	אימהו	2.049	אימהולר	1.216	אימהולרב	0.474
38	אימוה	3.658	אימוהלר	0.700	אימוהלרב	0.687
39	איהמו	1.736	איהמולר	1.674	איהמולרב	0.176
40	איהום	1.982	איהומלר	0.853	איהומלרב	0.123
41	איומה	1.628	איומהלר	1.280	איומהלרב	0.437
42	איוהמ	4.215	איוהמלר	0.774	איוהמלרב	0.263
43	ומהאי	3.127	ומהאילר	0.290	ומהאילרב	-
44	ומאהי	3.962	ומאהילר	1.049	ומאהילרב	-
45	והמאי	1.458	והמאילר	0.962	והמאילרב	0.448
46	והאמי	2.187	והאמילר	1.335	והאמילרב	0.184
47	והאימ	3.868	והאימלר	1.158	והאימלרב	-
48	והיאמ	3.339	והיאמלר	0.919	והיאמלרב	0.187
49	ואמהי	1.419	ואמהילר	1.288	ואמהילרב	-
50	ואהמי	8.261	ואהמילר	1.544	ואהמילרב	-

	A	Ω×100	B	Ω×100	C	Ω×100
51	ואהימ	6.225	ואהימלר	0.959	ואהימלרב	1.415
52	ואיהמ	5.457	ואיהמלר	0.736	ואיהמלרב	-
53	ויהאמ	15.151	ויהאמלר	0.736	ויהאמלרב	0.748
54	ויאהמ	2.739	ויאהמלר	0.741	ויאהמלרב	-
55	יהאומ	2.876	יהאומלר	0.682	יהאומלרב	-
56	יהואמ	3.518	יהואמלר	0.946	יהואמלרב	0.112
57	יאהומ	2.440	יאהומלר	1.076	יאהומלרב	0.277
58	יאוהמ	2.301	יאוהמלר	1.309	יאוהמלרב	0.253
59	יוהאמ	6.011	יוהאמלר	0.526	יוהאמלרב	0.125
60	יואהמ	13.772	יואהמלר	1.306	יואהמלרב	0.361
ממוצע		4.06		1.03		0.409

הטבלה ממחישה הבדל רציני וקבוע בין "מידת הקִרבה" של כל סוג וסוג:

- למעט 3 מקרים (שהובלטו ברקע אפור) מתוך 60, "מידת הקִרבה" של סוג A גדולה מזו של סוג B.
- למעט 3 מקרים (שהובלטו ברקע אפור) מתוך 45, בהם הוגדרה "מידת הקִרבה" של סוג C, "מידת הקִרבה" של סוג B גדולה יותר.
- בכל 45 המקרים, בהם הוגדרה "מידת הקִרבה" של סוג C, "מידת הקִרבה" של סוג A גדולה יותר.
- הממוצעים עבור כל סוג משקפים הבדל מהותי בין הסוגים.

כל "דרך סטטיסטית סטנדרטית" (כרצונו של דיאקוניס) תגלה, כי שלוש ההתפלגויות של ערכי "מידת הקִרבה", שנקבעו על ידי שלושת הסוגים הנ"ל, הן זרות זו לזו, ולכן ההנחה שהניח – שגויה לחלוטין.