

## טעות פשוטה

מאת דורון ויצטום

בקובץ "מדידת מפגשים" הגדרנו את "מידת הקרבה". זו מקבלת ערך גדול יותר אם המד"שים הנפגשים מינימליים יותר, קרובים יותר, זה לזה ו"לא מפורזים" במידה רבה יותר. שם גם הגדרנו את "מידת הקרבה המכילת" המודדת מה הסיכוי, של "מידת הקרבה" ערך כה גדול. כפי שהוסבר שם באריכות, "מידת הקרבה המכילת" 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 (ראה ב"צופן בראשית" פרק י'), כי לא הצליח להבין מדוע נזקקנו להגדיר את "מידת הקרבה המכילת". דיאקוניס חשב בטעות, כי ניתן להשוות ישירות את "מידת הקרבה" של זוג ביטויים אחד ל"מידת הקרבה" של זוג ביטויים אחר. מכאן הסיק, כי אפשר ליישם שיטות סטטיסטיות סטנדרטיות המבוססות על השוואה, ישירות על קבוצת ערכי "מידת הקרבה" 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 המכילת" מניבה מוזרויות ביחס להגדרת "מידת הקרבה". לדוגמא הוא מביא "פרדוקס":

"מידת הקרבה המכילת" 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 הקרבה המכילת" 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 הקרבה" מקבלים תמונה הפוכה: "מידת הקרבה" 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 מולדת, 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 המכויילת" 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 המכויילת" 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 לקרבה", שיוגדרו בקובץ "קירבה כוללת". אך לא נעסק בכך עכשיו, רק נדווח כי באותו מעמד אכן הוכחה טענת היוונים במובהקות ניכרת, ובטקס נעילת האולימפידה הוכרז על כך בחגיגות.

והנה, קם הנציג המצרי – בכיר הנציגים הזרים – וערער על התוצאות ועל המסקנות. לטענתו היה פגם בסיסי בהגדרת "מידת המהירות המכויילת" – 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 המכויילת" שלו היא  $50/125$ . לעומת זאת, בתחרות לבני 99 הגיע הרץ היווני למקום הראשון מתוך 101 רצים, כך ש"מידת המהירות המכויילת" 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 המכויילת", 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 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  $1700=4/6801$  טורים, נקבעה על ידי המד"ש המינימלי של המלה "הסתברות" בכל ספר בראשית:

\*\*\*\*\*

ברואתו יצחקו ישמעאלבן יואלמעריתהמכ  
 אמראהנהאשתכהואואיכאמרתאחתיהוא  
 בקלילאשראנימצוהאתכלכנאאלהצאנוק  
 ובלבו יקרבו ימיאבלאבי ואהרגהאתיעק  
 למקמהו יאמר להמייעקבאחי מאי נאתמו יא  
 ליהיעקבויתהרב לההותלדליעקבבנותאמ  
 שבימהפרידיעקבויתנפניהצאנאלעקדו  
 ספתהלביתאביכלמהגנבתאתאלהיויעני  
 מהלכלקראתכוארבעמאותאישעמווירא  
 שפחותהנהוילדיהנותרתחוניותגשגמל  
 אתבנתמנקהלנוולנשימואתבנותינונתל  
 ההפילגשאביווישמע ישראלויהיובני  
 אדומבלעבבעורושמעירודנהבהויתב  
 תואתכתנתהפסימאשרעליוויקהווישל  
 אלאתאנשימקמהלאמראההקדשהואבע  
 נחנוובעינישרביתהסיהרויתנשרביתהסה  
 יהיבבקרופעמרוחווישלחוויקראאתכל  
 צמצר ימויסרפרעהאתטבעתומעלדוית  
 משמרשלשת ימימויאמראלהמיוספביומה  
 קשוואמלאהביאתיואליוכהצגתיולפני  
 הרחיקוויספאמרלאשרעלביתוקומרדפ  
 יולאכלואחיולענותרתוכינבהלומפנ  
 ימהיעקבוכלזרעואתובניובניווא  
 עמדהולפניפרעהויברכיעקבאתפרעהו  
 לוישב עלהמתהוואמר יעקבאליוספאלשד

\*\*\*\*\*

על פני הטבלה מופיע מפגש קרוב ולא מפוזר עם מד"ש מינימלי של המלה המלאכותית "האמוי", בת 5 אותיות שכיחות (להלן יוסבר איך הגעתי אליה). למרות שהמפגש נראה "פוטוגני" – ואכן "מידת הקרבה" 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 המכויילת" של מפגש נאה זה (ראה בסוף הקובץ "מדידת מפגשים") 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  $680=10/6801$  טורים, נקבעה על ידי המד"ש המינימלי של המלה "הסתברות" בכל ספר בראשית:



לדוגמא, ביטוי בן 8 אותיות צפוי להופיע מספר פעמים מועט בדילוג שווה של אותיות, בדרך כלל רק פעם אחת בלבד (במקום כ-10 מד"שים המייצגים בדרך כלל את הביטוי). ולכן, במרוץ הנערך עבור זוג ביטויים, שאחד מהם בן 8 אותיות, ישתתפו מד"שים מועטים. נובע מכך, כי "מידת הקרבה" – שהיא סכום התרומות של מפגשי המד"שים – 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 הקרבה" במרוץ אחד אינה בת-השוואה לזו של מרוץ אחר. לפי טעותם, ערך התוחלת של "מידת הקרבה" 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 הקרבה" 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 יצרתי 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.

שים 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 הקרבה" במפגש עם המלה "הסתברות". לנוחות ההצגה הוכפל כל מספר פי 100. בטור האחרון סימנתי ב"- את המקרים שבהם הביטוי אינו מופיע בדילוג שווה.

$\Omega \times 100$	C	$\Omega \times 100$	B	$\Omega \times 100$	A	
0.325	אהוימלרב	1.118	אהוימלר	1.403	אהוימ	1
0.453	המאוימלרב	1.060	המאוימלר	1.810	המאוי	2
-	המאוימלרב	1.556	המאוימלר	2.483	המאוי	3
-	המאוימלרב	0.726	המאוימלר	6.842	המאוי	4
-	המאוימלרב	1.001	המאוימלר	3.517	המאוי	5
0.231	האמאוימלרב	0.663	האמאוימלר	3.100	האמאוי	6
-	האמאוימלרב	1.600	האמאוימלר	5.124	האמאוי	7
0.490	האמאוימלרב	0.842	האמאוימלר	6.678	האמאוי	8
0.626	האמאוימלרב	1.023	האמאוימלר	9.158	האמאוי	9
0.161	האימאוימלרב	1.144	האימאוימלר	4.038	האימאוי	10
0.447	האימאוימלרב	0.989	האימאוימלר	1.379	האימאוי	11
0.588	הומאוימלרב	0.702	הומאוימלר	1.732	הומאוי	12
0.400	הומאוימלרב	0.731	הומאוימלר	3.998	הומאוי	13
0.423	הואימלרב	1.635	הואימלר	0.601	הואימ	14
0.386	הואימלרב	1.035	הואימלר	3.506	הואימ	15
0.255	הימאוימלרב	1.077	הימאוימלר	4.193	הימאוי	16
0.295	הימאוימלרב	1.023	הימאוימלר	0.548	הימאוי	17
0.217	היאומלרב	1.215	היאומלר	3.260	היאומ	18
-	היאומלרב	1.006	היאומלר	1.987	היאומ	19
0.351	אמהוימלרב	0.850	אמהוימלר	1.608	אמהוי	20
0.562	אמהוימלרב	1.303	אמהוימלר	2.353	אמהוי	21
-	אמהוימלרב	1.461	אמהוימלר	1.488	אמהוי	22
0.402	אמאוימלרב	0.833	אמאוימלר	2.934	אמאוי	23
0.399	אמאוימלרב	1.392	אמאוימלר	10.600	אמאוי	24
0.106	אמאוימלרב	1.088	אמאוימלר	1.605	אמאוי	25
0.192	אהמאוימלרב	0.701	אהמאוימלר	1.898	אהמאוי	26
0.632	אהמאוימלרב	1.029	אהמאוימלר	10.518	אהמאוי	27
-	אהמאוימלרב	1.019	אהמאוימלר	2.775	אהמאוי	28
0.233	אהימאוימלרב	1.015	אהימאוימלר	1.403	אהימאוי	29
1.304	אהימאוימלרב	0.969	אהימאוימלר	4.816	אהימאוי	30
0.469	אומאוימלרב	0.541	אומאוימלר	24.714	אומאוי	31
0.399	אומאוימלרב	1.164	אומאוימלר	1.523	אומאוי	32
0.318	אוהמאוימלרב	1.215	אוהמאוימלר	2.781	אוהמאוי	33
0.750	אוהמאוימלרב	1.625	אוהמאוימלר	1.877	אוהמאוי	34
0.360	אוימאוימלרב	0.580	אוימאוימלר	1.071	אוימאוי	35
0.381	אוימאוימלרב	1.091	אוימאוימלר	0.906	אוימאוי	36
0.474	אימאוימלרב	1.216	אימאוימלר	2.049	אימאוי	37
0.687	אימאוימלרב	0.700	אימאוימלר	3.658	אימאוי	38
0.176	איהמאוימלרב	1.674	איהמאוימלר	1.736	איהמאוי	39
0.123	איהמאוימלרב	0.853	איהמאוימלר	1.982	איהמאוי	40
0.437	איומאוימלרב	1.280	איומאוימלר	1.628	איומאוי	41
0.263	איומאוימלרב	0.774	איומאוימלר	4.215	איומאוי	42
-	ומאוימלרב	0.290	ומאוימלר	3.127	ומאוי	43
-	ומאוימלרב	1.049	ומאוימלר	3.962	ומאוי	44
0.448	והמאוימלרב	0.962	והמאוימלר	1.458	והמאוי	45
0.184	והמאוימלרב	1.335	והמאוימלר	2.187	והמאוי	46
-	והמאוימלרב	1.158	והמאוימלר	3.868	והמאוי	47
0.187	והמאוימלרב	0.919	והמאוימלר	3.339	והמאוי	48
-	ואמאוימלרב	1.288	ואמאוימלר	1.419	ואמאוי	49
-	ואמאוימלרב	1.544	ואמאוימלר	8.261	ואמאוי	50
1.415	ואמאוימלרב	0.959	ואמאוימלר	6.225	ואמאוי	51
-	ואמאוימלרב	0.736	ואמאוימלר	5.457	ואמאוי	52

$\Omega \times 100$	C	$\Omega \times 100$	B	$\Omega \times 100$	A	
0.748	ויהאמלרב	0.736	ויהאמלר	15.151	ויהאמ	53
-	ויהאמלרב	0.741	ויהאמלר	2.739	ויהאמ	54
-	יהאומלרב	0.682	יהאומלר	2.876	יהאומ	55
0.112	יהואמלרב	0.946	יהואמלר	3.518	יהואמ	56
0.277	יההומלרב	1.076	יההומלר	2.440	יההומ	57
0.253	יאהמלרב	1.309	יאהמלר	2.301	יאהמ	58
0.125	יאהמלרב	0.526	יאהמלר	6.011	יאהמ	59
0.361	יאהמלרב	1.306	יאהמלר	13.772	יאהמ	60
0.409		1.03		4.06		ממוצע

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

- למעט 3 מקרים (שהובלטו ברקע אפור) מתוך 60, "מידת הקרבה" של סוג A גדולה מזו של סוג B.
  - למעט 3 מקרים (שהובלטו ברקע אפור) מתוך 45, בהם הוגדרה "מידת הקרבה" של סוג C, "מידת הקרבה" של סוג B גדולה יותר.
  - בכל 45 המקרים, בהם הוגדרה "מידת הקרבה" של סוג C, "מידת הקרבה" 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 הקרבה", שנקבעו על ידי שלושת הסוגים הנ"ל, הן זרות זו לזו, ולכן ההנחה שהניח – שגויה לחלוטין.