פרק עשרים

מבחן מכריע

שעת ההכרעה מִשמשה ובאה. לאחר התייעצויות נוספות עם עמיתיו הסטטיסטיקאים, ולאחר שדיאקוניס וסטטיסטיקאי מאוניברסיטת ברקלי נבחרו לשופטים רשמיים, זימן אותנו פרופסור אומן למשרדו ביום כ"ז חשוון התשנ"ב (4 בנובמבר 1991 למניינם). הוא ביקש כי נאשרר כמה תנאים לעריכת הניסוי. התנאי הראשון והעיקרי היה, כי נגיש את תוצאות הניסוי לפִרסום **בכל מקרה** – גם אם ניכשל בניסיון לעבור את "הרף" שיוצב בפנינו. בהתייחסותו של אומן לאפשרות זו, ניכר היה כי הושפע מאמונתם המוצקה של עמיתיו הסטטיסטיקאים, כי ניכשל בניסוי. נדמה לי, כי כוונתו בתנאי זה היתה לעורר אותנו למחשבה שניה על סיכויינו להצליח, ולאפשר לנו נסיגה בטרם קרב. אולם, אנחנו הסכמנו לתנאי זה.

לימים, נשאלנו מדוע באמת לא חששנו שמא ניכשל בניסוי, כפי שאכן ניבאה קבוצת הסטטיסטיקאים המפורסמים. סוף סוף, הרי מדובר בהערכתם של מיטב אנשי המדע בתחום זה, אנשים מיומנים ומשופשפים, שראו גם ראו תוצאות סטטיסטיות רבות, שעל פניהן נראו מרשימות – אבל קרסו במבחנים מחמירים כמבחן הראנדומיזציה.

התשובה היא, כי ידענו שאנשי מדע אלה מעולם לא חקרו את התופעה הנידונה לעומקה בחקירה חסרת פניות. הערכנו, כי לפי השקפת עולמם לא היה לסטטיסטיקאים אלה אף צלו של ספק באמונתם, שאין שום מידע מוצפן בספר בראשית. לכן היו בטוחים כי התוצאה הפנטסטית שהתקבלה בניסוי המקורי, אינה אלא השתקפות של תלות כזו או אחרת בין התוצאות עבור זוגות המדגם. לדעתם, מבחן הראנדומיזציה היה המבחן הנכון למקרה כזה, מבחן המסוגל להוכיח זאת ולהראות כי התוצאה הנכונה היא חוסר מובהקות של המדגם המקורי, כלומר – אפס גמור. לא היו לנו אשליות: כאשר הנושא הוא "הטיה בשיפוט עקב דעות קדומות" – גם לאנשי מדע מפורסמים אין שום יתרון ביחס לכל אדם אחר. בניגוד להם, ממחקרנו בספר בראשית היה לנו ביטחון מלא, כי התופעה אמיתית, ולכן, התוצאה הכוללת שנתקבלה בניסוי המקורי נובעת מן העובדה, כי תאריכי הלידה והפטירה של האישים הנתונים אכן הוצפנו בכוונה תחילה בספר בראשית. דווקא האפשרות שהתוצאה בניסוי המקורי היא תולדה של תלות מסתורית – נראתה בעינינו כבלתי סבירה במידה קיצונית.

בצד התחייבותנו להגיש את תוצאות הניסוי לפרסום בכל מקרה, התחייב אומן לשלוח את התוצאות ל- *PNAS* במקרה של הצלחה. אולם, הוא הודיע לנו, כי עדיין לא החליט מהו הסף

להצלחה (אומנם הוא עתיד לקבוע זאת בהמשך היום), וכי על פי עצת אחד השופטים, הוא יודיע מהו הסף רק לאחר סיום הניסוי.

פרופסור אומן גם השאיר בידיו את "המפתח" "להתנעת" הניסוי. הכוונה – למספר שיש להזין לתוכנית המחשב, כדי שתבָחר סדרה **מסוימת** של 999,999 צימודים פסודו-רנדומליים לצורך הניסוי (פרטים נוספים בנספח א7). מספר זה, הנקרא "זרע סטטיסטי", היה מספר בן 32 ספרות בינאריות (כלומר, מספר שלם בעל 32 ספרות, כאשר הוא נכתב לפי בסיס 2). הוא נקבע יום קודם לכן, באמצעות כמה שיחות טרנס-אטלנטיות, באופן הבא. אומן התקשר לשלושה מחמשת הסטטיסטיקאים, וביקש מכל אחד מהם לספק מספר כזה. דיאקוניס יצר מספר כזה על ידי 32 הטלות מטבע, ואילו השניים האחרים השתמשו בזוגיות הספרות בפיתוח העשרוני של המספר π, בשני אזורים רחוקים זה מזה. שלושת המספרים חוברו על ידי אומן[1], אשר מסר לנו את "הזרע"[2]

01001 10000 10011 11100 00101 00111 11

(בהצגתו הרגילה, העשרונית, המספר הוא 1,277,674,143). אומן איחל לנו הצלחה, ובזה התחיל הניסוי, מבחינתו.

--- --- ---

מבחינתנו, החל הניסוי זמן מה אחר כך, עם השלמת כמה הכנות טכניות. הניסוי, שהחל כמעט שש שנים לאחר התחלת הניסוי הגדול על המדגם הראשון, היה שונה ממנו מבחינות רבות.

מצד אחד, היקף החישובים ומורכבות הפעולות היו גדולים בהרבה: במסגרת מחקרנו, מעולם לא ביצענו עבודה בסדר גודל כזה. מצד שני, האמצעים שעמדו לרשותנו היו משוכללים בהרבה. מחשב ה- 386 בעל 20 מגה-הרץ בצירוף התוכנה המשופרת שהכין יואב, אִפשרו מהירות ביצוע "דמיונית"[3] – כך שכל נפח החישובים העצום הִצריך פחות זמן מן הניסוי המקורי. בשלב הראשון, הכינה התוכנה של יואב את המאגר הגדול של הנתונים: תוצאות המפגשים של הכינויים במדגם עם כל התאריכים במדגם. בשלב השני, הרצתי תוכנה אחרת, אשר הוכנה על ידי יעקב רוזנברג, אשר הגרילה את 999,999 הצימודים האקראיים (כמבואר בנספח א7), "שלפה" את הנתונים המתאימים להם ממאגר הנתונים וערכה תחרות בין מידות "הנטייה הכוללת לקִרבה" של 999,999 המדגמים המשובשים, שנוצרו בדרך זו, לבין המדגם המקורי. יעקב, שהוא מתכנת מעולה (הוא אינו קרוב משפחה של יואב), השקיע מאמצים ניכרים וכשרון רב כדי ליצור תוכנית יעילה ומהירה שאִפשרה את ביצוע הניסוי בזמן סביר.

במהלך הניסוי ערכנו ארבע תחרויות במתכונת מבחן הראנדומיזציה (שהוסבר בפרק הקודם), עבור ארבע[4] מידות של "הנטייה הכוללת לקִרבה" שנקבעו מראש (ראה בנספח א3). בכל תחרות השתתפו אותם 999,999 מדגמים משובשים. הדירוגים של ארבע המידות מובאים בנספח א8 (בתוספת פרטים והסברים). אחת המידות, זכתה במקום **הרביעי** בדירוג מתוך 1,000,000 מתחרים.

---

[1] מודולו 2^{32}.

[2] לא לפני שהועמד בפנינו תנאי נוסף: אם נצליח בניסוי ונתבקש לערוך אותו מחדש באופן אחר (ראה נספח א7) – עלינו להסכים לכך.

[3] בזמן כתיבת שורות אלו יש שימוש נרחב במחשבים אישיים המהירים פי מאות.

[4] אלה אותם שני סטטיסטים P_1 ו- P_2 שהוזכרו בפרק י, אשר חושבו פעם אחת עבור המדגם השני כולו, ופעם אחרת עבור המדגם השני החלקי – מדגם ב1 שהוגדר בפרק יז.

כלומר, היו רק 3 מדגמים מתוך 999,999 המדגמים המשובשים, שערך מידת "הנטייה הכוללת לקרבה" שלהם היה **קטן** מזה של המדגם המקורי.

ההסתברות שהדירוג הוא כה טוב – כלומר: ההסתברות, שערכה של מידת "הנטייה הכוללת לקִרבה" של המדגם המקורי כה נמוך – הוא: $p = 0.000004$ (אחד למאתיים וחמישים אלף).

כדי לקבל את המובהקות הסטטיסטית של תוצאות הניסוי, יש לחשב את ההסתברות הכוללת לתוצאה כזו. בחישוב כזה מתחשבים בעובדה שנערכו בסך הכל ארבעה מירוצים עבור ארבע מידות. פרופסור אומן ערך את החישוב בדרך שנקבעה מראש[5], וקיבל כי

המובהקות הסטטיסטית של תוצאות הניסוי היא $p = 0.000016$.

מובהקות כזו (הטובה מאחד לששים אלף) נחשבת למובהקות פנטסטית בהשוואה לנדרש בניסויים מדעיים. רק בשלב זה גילה לנו אומן, כי הוא הציב את "רף" המובהקות על הערך החריג של $p = 0.0033$ (אחד לשלוש מאות). אכן, "ניתרנו" הרבה למעלה מן "הרף", אשר נותר אֵי שם מתחתנו.

כאשר התעשת פרופסור אומן מן ההפתעה, הוא ביקש לבדוק את תקינות הניסוי באמצעות הרצת הניסוי על טקסט אחר לביקורת. חזרנו על הניסוי עם "טקסט R" שהתקבל מִספר בראשית אחרי שעִרבבנו את אותיותיו (ראה נספח א7). התוצאות היו בלתי מובהקות לחלוטין (פירוט התוצאות בנספח א8).

שוב ניהל פרופסור אומן שיחות טרנס-אטלנטיות עם עמיתיו הסטטיסטיקאים. הפעם דיווח להם על התוצאות. התדהמה מעבר לקו היתה מושלמת. אחד מהם, איש מדע מפורסם וחתן פרס נובל[6], מִלמל דבר מה על "מלחמה" ועל "שלום". תחילה, אומן לא הבין: איזו מלחמה? מלחמת המפרץ? – זו הרי נסתיימה כבר כמה חודשים קודם לכן! אך עד מהרה הובהר לו, שעמיתו מבקש לערוך בדיקה נוספת: לחזור להריץ את הניסוי – הפעם על טקסט נוסף – על הספר "מלחמה ושלום" של טולסטוי...

"מדוע?" – תמהתי, "האומנם איש המדע הנכבד אינו מאמין בחוקי הסטטיסטיקה?"

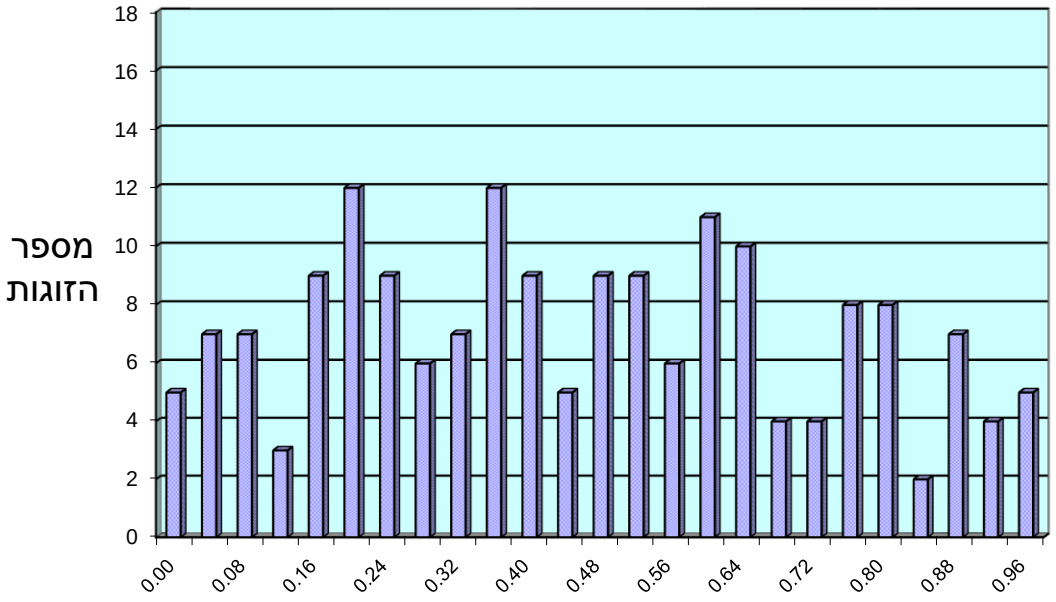
לא ציפיתי לתשובה. השגתי את ספרו של טולסטוי, "מלחמה ושלום", מתורגם לעברית בידי לאה גולדברג. היה צורך להקליד ולהגיה את חלקו הראשון של "מלחמה ושלום", קטע בן 78,064 אותיות (בדיוק כאורך ספר בראשית). אחר כך היה עלי להריץ את כל הניסוי מחדש עם "טקסט T" (כך כינינו אותו). כל העניין היה מיותר, חשתי כי אני טוחן מים במחשב משוכלל. בעיקר היה חבל על הזמן: עם כל מאמצי להחיש את הדברים, הכנת הטקסט והרצת המחשב גזלו כחודש ימים. רק ב- 7 בינואר 1992 (למניינם) היו בידי התוצאות. כצפוי, התוצאות היו בלתי מובהקות לחלוטין. פירוט התוצאות בניסוי ניתן בנספח א8, כאן אציג רק היסטוגרמה של תוצאות המדגם המקורי ב"מלחמה ושלום".

---

[5] הוא השתמש באי השוויון של בונפרוני (Bonferroni). אומנם, השימוש בו במקרה זה מחמיר מדי (ראה נספח א8).

[6] Kenneth Joseph Arrow.

המדגם השני: התפלגות התוצאות בטקסט T



ערכי "מידת הקרבה המכוילת"

איור כ-1

בעצם, מבחן הראנדומיזציה על "מלחמה ושלום" רק אישר את ההתרשמות המיידית מאִיור זה: תוצאות מפגשי זוגות המלים מתפלגות באופן אקראי.

--- --- ---

תוצאות מבחן הראנדומיזציה על ספר בראשית היוו הפתעה לא רק לפרופסור אומן ולשופטים; אף אנו הופתענו – אך דווקא מן הסיבה ההפוכה. אכן, ציפינו לתוצאה טובה יותר. כדי למנוע אי-הבנה אבהיר: המובהקות שהושגה בניסוי היתה יוצאת מן הכלל – פנטסטית לפי אמות מידה מדעיות מקובלות, ומספיקה בהחלט לפרסום הניסוי. לא היתה לנו שום סיבה להתלונן. אלא שנתגלע פער ניכר בינה לבין ההערכה המקורית של הצלחת המדגם השני באמצעות מידות "הנטייה הכוללת לקִרבה", כך שהיה עלינו לדעת מה מקורו של פער זה.

אומנם, אז לא ידענו כי הפער קטן מכפי שהוא נראה. רק כעבור זמן, כאשר גם חוקרים וגם מתנגדים ברחבי העולם הריצו מחדש את הניסוי, ואף ניסו זאת בהשוואה לקבוצה גדולה יותר של מדגמים משובשים, התברר כי התוצאה טובה יותר. כעבור שנים, אף אני הרצתי את הניסוי מחדש, כאשר מידת "הנטייה הכוללת לקרבה" של המדגם המקורי מתחרָה ב- 199,999,999 מדגמים משובשים. התוצאה עבור המידה "המוצלחת" השתפרה בערך פי 6:

מ- $p = 0.000004$ ל- $p = 0.00000066$,

כך שהמובהקות הכוללת המדויקת יותר עבור הניסוי היא:

$p = 4 \times 0.00000066 = 0.00000264$ (בערך אחד לשלוש מאות ושמונים אלף).

למרות זאת, עדיין נותר פער ניכר בין התוצאה בניסוי לבין הערכתנו המקורית. הפער המזערי הוא עבור המידה "המוצלחת": פקטור של 100 (בערך). עבור המידות האחרות הפער היה גדול בהרבה.

במחשבה ראשונה סברנו, כי שגיאה בהערכתנו המקורית גרמה לפער זה. כבר הזכרתי בפרק יח, כי דיאקוניס טען שעשינו שימוש בשתי הנחות בלתי מוצדקות[7]. מן הניסוי התברר לכאורה, כי הוא צדק בטענה זו. רק לאחר זמן רב התברר, כי יש כאן גורם נוסף. חלק מן הפער הנ"ל נוצר כנראה מסיבות אחרות לגמרי (עליהן נעמוד בנספח א10). בכל מקרה, התברר מן הניסוי, כי לשגיאה זו היתה השפעה מוגבלת בלבד על התוצאה המקורית, ובזה אכן טעה דיאקוניס[8] בגדול: התוצאה המקורית לא היתה תולדה של שגיאה זו.

לסיכום: מבחן הראנדומיזציה הוכיח, כי התוצאה המקורית היתה תולדה של הצלחת המפגשים של כינויי האישים ותאריכיהם, וכי אכן קיימת כאן מובהקות סטטיסטית מדהימה. אלא, שההערכה המקורית של גודל ההצלחה היתה שונה מן התוצאה במבחן הראנדומיזציה – בגלל סיבות שונות (עיינו בנספח א10).

---

[7] כפי שמבואר בנספח א3, הנחנו לצורך חישוב התוצאה הכוללת, שערכי "מידת הקרבה המכוילת" הם בלתי תלויים ומתפלגים בצורה אחידה (אוניפורמית). במבחן הראנדומיזציה לא נעשה שימוש בהנחות אלה.

[8] דיאקוניס טעה גם בהשערתו על הסיבות שגרמו לתלות בין ערכי "מידת הקרבה המכויילת". סיבות אלו נחקרו בשנים שלאחר מכן והדיון בהן ייערך במקום אחר.