בס"ד טיוטה אחרונה: ח' תשרי התשע"ט

ניסוי חדש בתחום המורחב

מאת דורון ויצטום

תקציר

המאמר "*Equidistant letter sequences in the Book of Genesis*" [1] מאת דורון ויצטום, אליהו ריפס וויואב רוזנברג (להלן ור"ר), אשר פורסם בכתב העת Statistical Science בשנת התשנ"ד, הציג תוצאות מחקר מדעי של הרמז בדילוג השווה בספר בראשית. מחקר זה עסק בקבוצה של מופעי ביטויים בדילוגים שווים, שהם מינימליים בקטעים גדולים בטקסט הנחקר. מופעים כאלה נכנה כאן "צפני *ELS*". במוקד המחקר היה ניסוי שבו נמדדה הקורלציה בין צפני *ELS* של השמות והכינויים של קבוצת אישים (גדולי תורה) לבין צפני *ELS* של תאריכי הלידה והפטירה שלהם. המדידה הראתה "סיגנאל" ברור. המובהקות הסטטיסטית של הסיגנאל נמדדה באמצעות מבחן פרמוטציות, דהיינו, רנדומיזציה של הנתונים. התוצאה היתה מובהקת ביותר, ברמה של 0.00002. תאור מפורט בעברית של העבודה פורסם בבד"ד [2].

במאמרנו "סיגנאל של מערכת צפני *ELS*: דעיכה מול הגברה" [3] נמדדה הקורלציה בין צפני *ELS* עבור הנתונים הנ"ל בתחום מדידה מורחב. תחום המדידה המורחב הינו קטע מתחילת ספר התורה, הכפול באורכו מתחום המדידה המקורי, שהיה ספר בראשית בלבד. התברר, כי בתחום המדידה המורחב הסיגנאל מתחזק באופן בולט ומובהק. מסקנת המאמר היתה, כי תחום ההצפנה של הנתונים הנ"ל קרוב יותר לתחום המורחב, מאשר לתחום המקורי. במאמרנו "סיגנאל של מערכת צפני *ELS*: ניסויים חדשים ב"תחום המורחב" [4] דווח על ניסויים נוספים בנושא רשימת הנתונים הנ"ל בתחום המורחב. בכולם נתקבלה מובהקות חזקה. בעבודות [3] ו[4] נמדדה המובהקות באמצעות רנדומיזציה של הטקסט.

בעקבות העבודות [3] ו[4], נעשה כאן עוד ניסוי בנושא רשימת הנתונים הנ"ל בתחום המורחב. הפעם, הניסוי מבוסס על רשימת נתונים, ששימשה את מבקרי המחקר בניסוי העיקרי מתוך שלושה ניסויים "החשובים ביותר" ("Perhaps the most important class of experiments we have conducted") [5] (עמ' 164-163). רשימת נתונים זו, שנועדה לחקות את רשימת הנתונים ששימשה בניסוי [1] של ור"ר, הוכנה לדבריהם על ידי מומחה בלתי תלוי, שמחה עמנואל. הם מדווחים כי תוצאות הניסוי על רשימת "החיקוי" בספר בראשית אינן מובהקות במבחן הפרמוטציות, והם רואים בכך ראייה מכרעת לטענתם המרכזית כי הצלחת הניסוי של ור"ר נבעה אך ורק ממניפולציה של הנתונים.

נעשו כאן כמה בדיקות.

בשלב ראשון נבדק: האם קיים סיגנאל לא-טריוויאלי של רשימת "החיקוי" בספר בראשית?
התוצאות בספר בראשית היו מובהקות בהחלט. באחת המדידות נתקבלה מובהקות של $p = 7.8E{-4}$.
המובהקות הכוללת של התוצאות בספר בראשית היא $p = 1.38E{-3}$.

בשלב שני נבדק: האם הסיגנאל של רשימת "החיקוי" מתגבר או דועך במעבר לתחום המורחב?
התוצאה היתה של התגברות הסיגנאל.

בשלב השלישי נבדקה מובהקות הסיגנאל של רשימת "החיקוי" בתחום המורחב.
התוצאות בתחום המורחב היו מובהקות ביותר. באחת המדידות נתקבלה מובהקות של $p < 2E{-5}$.
המובהקות הכוללת של התוצאות בתחום המורחב היא $p = 1E{-4}$.
כמו בעבודות [3] ו[4], גם כאן כל המובהקויות נקבעו באמצעות רנדומיזציה של הטקסט.

אכן, כפי שכתבו מתנגדי המחקר, לרשימת "החיקוי" שנוצרה על ידי מומחה מטעמם יש חשיבות גדולה מאד: התוצאות דלעיל מפריכות את טענתם המרכזית הנ"ל. הן מוכיחות, כי יש בתורה תופעה אמיתית של הצפנת נתונים, וכי התוצאה המקורית של ור"ר איננה תולדה של מניפולציה של שמות וכינויים.

מבוא

א. הגדרות:

1. יהא T טקסט, נגדיר קטע בן n אותיות D, בטקסט T כך:

$$D \equiv ([t_{d_1}, t_{d_n}], T)$$

כאשר t_{d_1} הוא המספר הסידורי ב- T של d_1, האות הראשונה בקטע D, ו-t_{d_n} הוא המספר הסידורי ב- T של d_n, האות האחרונה בקטע D.

2. יהא T טקסט, D קטע ממנו, $LIST$ רשימה של זוגות ביטויים, ו- $ELS(LIST, D)$ קבוצת צפני ELS המייצגים את הביטויים ב- $LIST$ והמוגדרים ב- D. בניסוי המקורי של ור"ר ([1], עמ' 435-434; [2], עמ' 61-60), הוגדרה "מידת הקירבה המכוילת" $c(w, w')$ עבור כל זוג ביטויים (w, w'). "מידת הקירבה המכוילת" בין הביטויים, $c(w, w')$, היא מספר בין 0 ל- 1: ערכו קרוב ל- 0 כאשר צפני ELS של הביטויים (w, w') נפגשים בצורה מכונסת במיוחד; ערכו קרוב ל-1 כאשר צפני ELS של הביטויים (w, w') רחוקים זה מזה ומפוזרים במיוחד.

3. כדי לסכם את "הנטייה הכוללת לקירבה" עבור כלל הזוגות ב- $LIST$, הוגדרו בניסוי המקורי של ור"ר ([1], עמ' 436) שתי "מידות לנטייה הכוללת לקירבה" (המשמשות כסטטיסטים) באמצעות שני אופרטורים $\boldsymbol{P_1}$ ו- $\boldsymbol{P_2}$, המיושמים לגבי הטקסט, הקטע והרשימה הנתונים. המידות המתקבלות כאן יהיו:

$$P_i \equiv \boldsymbol{P_i}(LIST, D), \quad i = 1, 2.$$

הגדרה מתמטית מלאה של שני הסטטיסטים נמצאת בנספח, סעיף א. שם מבואר, כי כל אחד מהסטטיסטים P_i מקבל ערך הקטֵן ככל שהמפגשים בין צפני ה- ELS יותר "מוצלחים". כאן רק נעיר כי 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 $c(w, w')$.

4. נגדיר סיגנאל S_i באופן הבא:

$$S_i(LIST, D) \equiv 1 / P_i(LIST, D), \quad i = 1, 2.$$

לפי האמור בסעיף הקודם, הסיגנאל יקבל ערך הגדל ככל שהמפגשים בין צפני ה- ELS יותר "מוצלחים". (הגדרת הסיגנאלים S_i אינה מחדשת דבר מהותי לעומת ההגדרה המקורית של הסטטיסטים P_i; אומנם, היא נועדה לשמר את המינוח שבמאמרים הקודמים, [3] ו[4], אשר מאמר זה מבוסס עליהם.)

5. יהא D' קטע ב- T, ארוך מ-D והמכיל אותו. הסיגנאלים המתקבלים עבור $LIST$ ב- D' יסומנו $S_i(LIST, D')$. נגדיר את מדד ההגברה Q_i: $Q_i \equiv S_i(LIST, D') / S_i(LIST, D)$

$Q_i < 1$ מצביע על דעיכה, היחלשות S_i, ואילו $Q_i > 1$ מצביע על הגברת S_i.

יתכן כי עבור i שונים תתקבל מידת הגברה שונה. אף ייתכן ש- Q_1 יצביע על הגברה, בעוד Q_2 מצביע על דעיכה, או להיפך. נגדיר את מדד ההגברה המירבית Q:

$$Q \equiv \max\{Q_1, Q_2\}$$

ב. הרנדומיזציה:

את מובהקות הסיגנאל, כלומר, הסיכוי "לקבל סיגנאל כה גבוה במקרה", ניתן לקבוע באמצעות מבחן רנדומיזציה. בשנים האחרונות, עם השיפור הרב בביצועי המחשב, נתאפשר להשתמש במבחן רנדומיזציה של הטקסט. מבחן זה עדיף על מבחן הפרמוטציות בו נעשה שימוש ב[1], כפי שנתבאר בנספח, סעיף ב. אומנם, מבחן זה מצריך נפח חישובים גדול בהרבה, ולכן לא היה מעשי בעת ביצוע עבודת ור"ר [1].

1. במכתב לדוד קשדן [6] הציע אליהו ריפס את העקרונות הבאים לגבי מדידת מובהקות של רשימה בספר בראשית:

- להשתמש ב- 1000 (או יותר) טקסטים T_j, כל T_j נוצר מספר בראשית ע"י ערבוב אקראי של מלים בתוך פסוקים (הדבר נעשה בדיוק כפי שנעשה ב[1] לגבי טקסט U).

- להעריך את מובהקות הסטטיסטים P_i בספר בראשית, באמצעות השוואתם לערכי P_i^j המתקבלים בטקסטים T_j.

מבחן מסוג זה, המבוסס על רנדומיזציה של הטקסט הנחקר, יכונה להלן "מבחן טקסטים"; נפרט:

(א) מודדים את ערכי הסטטיסטים P_i עבור המדגם הנחקר.

(ב) חוזרים על המדידות עבור N טקסטים מעורבבים T_j (כל טקסט T_j נוצר מהטקסט הנחקר כנ"ל).

(ג) מוצאים את n_i, מספר הטקסטים T_j שלהם ערך $P_i^j \le P_i$.

(ד) הסיכוי לקבל ערך P_i כל כך קטן במקרה הוא $p_i = {n_i}/{N}$.

2. בכל העבודות בהן חושבה המובהקות באמצעות מבחן הטקסטים, נעשה שימוש בקבוצה של 50,000 טקסטים מעורבבים כנ"ל לכל היותר. (הגבלה זו נובעת מיכולות החישוב העומדות לרשותי.)

ג. רשימת הנתונים בניסוי שלפנינו:

1. רשימת הנתונים שנמדדה בעבודת ור"ר [1] בנויה מזוגות ביטויים (w, w'). בכל זוג, w הוא שם אישיות (או כינויה) מקבוצת אישים (גדולי תורה) ו-w' - תאריך הלידה או הפטירה שלה. השמות והכינויים נקבעו על ידי מומחה חיצוני, שלמה זלמן הבלין מאוניברסיטת בר-אילן.

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[1].

3. לכן, ניתן להציג את רשימת הנתונים השלמה, *LIST 2*, כאיחוד של שתי תת-רשימות:

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

$$LIST\ 2 = \text{הרשימה הכוללת} = L2 + L'2$$

אבחנה זו נעשתה לפני ביצוע מבחן הפרמוטציות ע"י ור"ר [1] (עמ' 436 בסופו – 437). מבחן הפרמוטציות, שפרטיו סוכמו מראש בין ישראל אומן לפרסי דיאקוניס, נועד למדוד את מובהקותם של ארבעה סטטיסטים: ערכי $\boldsymbol{P_1}$ ו-$\boldsymbol{P_2}$ לגבי הרשימה כולה *LIST 2* וערכיהם לגבי *L2*.

4. המובהקות החזקה ביותר, 4 למיליון, נתקבלה עבור הסטטיסטי P_2 עבור תת-הרשימה *L2*, וכך נקבעה הצלחת הניסוי.

5. טענתם המרכזית של מבקרי המחקר ברנדן מק-קי, דרור בר-נתן, מיה בר-הלל וגיל קלעי (להלן מבב"ק) במאמרם [5] היא, כי תת-הרשימה *L2* ניתנת "לתפירה" ע"י בחירה/השמטה מכוונת של שמות וכינויים (בניגוד לתת-הרשימה *L'2* הבנוייה מכינויים הסטנדרטיים רבי פלוני שהם קבועים). לכן, כאשר הכינו באופן מוצהר רשימה "תפורה" של נתונים כדי שתצליח בספר "מלחמה ושלום" [7], רק תת-הרשימה ללא הכינויים הסטנדרטיים תרמה להצלחה.

6. לפי [5] (עמ' 163-164), הכין שמחה עמנואל, מומחה בהיסטוריה רבנית באוניברסיטת תל-אביב, שלוש רשימות של שמות וכינויים של אישים (גדולי תורה) עבור מבב"ק לצורך מדידות צפני *ELS* בספר בראשית. העבודה נעשתה בהדרכתם של מבב"ק וללא ידיעת ור"ר. לדבריהם, יתכן שאלה הניסויים החשובים ביותר שהם עשו[2]. בניסוי העיקרי הוכנה רשימה, שנועדה על ידי מבב"ק "לחקות" את *LIST 2*, ונכנה אותה כאן *LIST E2*. רשימה זו מכילה שמות וכינויים של האישים הכלולים ב-*LIST 2* אשר נאספו על ידי עמנואל מבלי שראה את השמות והכינויים שקבע הבלין עבור *LIST 2*. גם את הרשימה החדשה ניתן להציג כאיחוד של שתי תת-רשימות:

- תת-הרשימה *E2* הבנויה מאותם זוגות ביטויים (w, w'), שבהם w הוא מן השמות והכינויים המיוחדים.
- תת-הרשימה *E'2* הבנויה מאותם זוגות ביטויים (w, w'), שבהם w הוא הכינוי רבי פלוני.

---

[1] כינויים סטנדרטים אלה נמצאים בעמודה מיוחדת בטבלת השמות והכינויים שהכין הבלין (ראו בבד"ד [2] עמ' 73-74).

[2] בלשונם: "Perhaps the most important class of experiments we have conducted".

$LIST\ E2$ = הרשימה הכוללת = $E2 + E'2$

7. גם במדידת הרשימה החדשה ע"י מבב"ק נתקבלה המובהקות המירבית במבחן הפרמוטציות באמצעות הסטטיסטי P_2 עבור תת-הרשימה $E2$. הערך בו הם נקבו [5] הוא 0.017.

8. רשימת הנתונים בניסוי שלפנינו:
אלה הם נתוני תת-הרשימה $E2$ בדיוק כפי שפורסמו על ידי מבב"ק (למעשה, הם פירסמו את $LIST\ E2$, שהעתק ממנה נמצא בנספח, סעיף ג).

ד. מבנה המדגם:
המדגם נוצר מרשימת הנתונים. A_i היא קבוצת השמות והכינויים השייכים לאישיות i, ו-B_i היא קבוצת ביטויי תאריכי הלידה ו/או הפטירה השייכים לאישיות i. המדגם כולל את כל זוגות הביטויים (w,w') כך ש-w הוא שם או כינוי ב- A_i ו- w' הוא תאריך (לידה ו/או פטירה) ב- B_i, לכל $1 \le i \le n$.

ה. התחום המורחב:
1. תחום המדידה של הניסוי המקורי של ור"ר [1] היה ספר בראשית.
2. במאמרנו "סיגנאל של מערכת צפני ELS: דעיכה מול הגברה" [3] נמדדה הקורלציה בין צפני ELS עבור תת-הרשימה $L2$ בתחום מדידה מורחב. תחום המדידה המורחב הוא קטע מתחילת ספר התורה, הכפול באורכו מתחום המדידה המקורי. התברר, כי בתחום המדידה המורחב הסיגנאל מתחזק באופן בולט ומובהק במקום לדעוך. מסקנת המאמר היתה, כי תחום ההצפנה של $L2$ קרוב יותר לתחום המורחב, מאשר לתחום המקורי (ספר בראשית בלבד).
3. במאמרנו "סיגנאל של מערכת צפני ELS: ניסויים חדשים בתחום המורחב" [4] דווח על ניסויים נוספים בנושא רשימת הנתונים $L2$ בתחום המורחב. בכולם נתקבלה מובהקות חזקה.
4. עבודות [3] ו[4] מוכיחות כי התוצאה המקורית של ור"ר איננה תולדה של מניפולציה של שמות וכינויים כפי שטוענים מבב"ק. (כפי שהוסבר והודגם ב[3], סיגנאל של רשימת נתונים הנוצר באופן מלאכותי בספר בראשית באמצעות בחירה/השמטה מכוונת של שמות וכינויים, דועך באופן חד כאשר בודקים אותה רשימה בתחום המורחב. ההפך מזה נצפה עבור $L2$.)
5. כפי שנתבאר בסעיף ג, תת-הרשימה $E2$ נוצרה על מנת "לחקות" את $L2$. האם גם הסיגנאל של רשימת "החיקוי" מתחזק בתחום המורחב? האם הוא מובהק? – כדי לענות על שאלות אלו, נמדוד את המדגם הנוצר מתת-הרשימה $E2$ הן בספר בראשית והן בתחום המורחב.

מדידת תת-הרשימה $E2$ בספר בראשית ובתחום המורחב

מבב"ק פירסמו רשימת נתונים של 33 אישים בשמו של שמחה עמנואל. אולם, כבר לפני זמן רב התברר [8], כי לאמיתו של דבר עמנואל הכין לבקשת מבב"ק רשימה גדולה יותר של 35 אישים, מהם 32 האישים שהיו אישי $LIST\ L2$ של ור"ר, ועוד שלושה אישים. מבב"ק השמיטו מרשימת עמנואל המקורית שני אישים הכלולים ב-$LIST\ L2$, הותירו את שלושת האישים הנוספים, ופירסמו את הרשימה "המשופצת" כרשימת עמנואל (המסומנת כאן $LIST\ E2$ ומובאת בנספח סעיף ג). מלבד זאת התברר, כי מבב"ק התערבו מבחינות נוספות בהכנת הרשימה [8].

למרות זאת, בדקנו את תת-הרשימה $E2$ כפי שהיא הן בספר בראשית והן בתחום המורחב. במדידות שלפנינו, נמדדו שלושה מדגמים שנוצרו מתת-הרשימה $E2$. ההבדל בין המדגמים הוא אך ורק בתצורות התאריכים בהן נעשה שימוש, כפי שיפורט לקמן.

א. המדגמים ששימשו לניסויי ור"ר [1] הן רשימות של זוגות ביטויים. בכל זוג, אחד מבני הזוג הוא שם או כינוי של אישיות מסוימת, ובן זוגו הוא תאריך הלידה או הפטירה של אישיות זו. כל תאריך הוצג בשלוש תצורות המסומנות ב-A, ב-B וב-C. למשל היום הראשון בחדש תשרי הוצג כך:
A - "א' תשרי", B - "בא' תשרי", C - "א' בתשרי".
הדבר נקבע באופן אפריורי לפני הכנת רשימת הנתונים הראשונה. בעקבות רעיון של מבב"ק (העניין מבואר היטב ב[4]) נערכו שתי מדידות נוספות עבור $L2$:

- כאשר משתמשים גם בתצורה רביעית: D – "בא' בתשרי". כלומר, ננקטות התצורות $ABCD$.

- כאשר משתמשים רק בשתי התצורות *AB*.

כך נעשה ב[4] עבור רשימות *L2* ו- *L2M*. כך נעשה ב[9] עבור רשימות אחרות.

ב. נגדיר עבור תת-הרשימה *E2* את המדגמים הבאים:

מדגם [א]: *E2(AB)* הוא המדגם הנוצר מתת-הרשימה *E2* כאשר התאריכים בה ניתנים בצמד התצורות AB.

מדגם [ב]: *E2(ABC)* הוא המדגם הנוצר מתת-הרשימה *E2* כאשר התאריכים בה ניתנים בשלש התצורות ABC.

מדגם [ג]: *E2(ABCD)* הוא המדגם הנוצר מתת-הרשימה *E2* כאשר התאריכים בה ניתנים בארבע התצורות ABCD.

ג. כאן יוצגו כמה מדידות.

במדידה 1, נמדד הסיגנאל S_2 עבור כל אחד מהמדגמים [א], [ב] ו[ג] בספר בראשית. מובהקות הסיגנאל נמדדה באמצעות מבחן הטקסטים.

במדידה 2, נערכו מדידות דומות בקטע מספר "מלחמה ושלום", שאורכו כאורך ספר בראשית.

במדידה 3, נמדדו הסיגנאלים עבור כל אחד מהמדגמים [א], [ב] ו[ג] בתחום המורחב. מובהקות הסיגנאלים נמדדה באמצעות מבחן הטקסטים.

במדידה 4, נערכה מדידה דומה למדידה 3, אך בקטע מספר "מלחמה ושלום", שאורכו כאורך התחום המורחב.

מדידה 1: מדידת הסיגנאל S_2 עבור תת-הרשימה *E2* בספר בראשית.

א. מה נמדוד:

1. כפי שנכתב במבוא (סעיף ג), ניתן להציג את רשימת הנתונים של ור"ר, *LIST 2*, כאיחוד של שתי תת-רשימות: תת-הרשימה *L2* הבנויה מאותם זוגות ביטויים (w, w'), שבהם w הוא מן השמות והכינויים המיוחדים, ותת-הרשימה *L'2* הבנויה מאותם זוגות ביטויים (w, w'), שבהם w הוא הכינוי רבי פלוני.
2. בניסוי של ור"ר [1] התברר, כי ההצלחה במבחן הפרמוטציות נקבעה ע"י *L2*, וכי הוספת *L'2* לרשימה הכוללת רק גרעה מן ההצלחה. (לפי מדידות שערכו מבב"ק בעצמם [5], מובהקות *L2* חזקה פי שלוש ממובהקות *LIST 2*.)
3. והנה, (כפי שנכתב במבוא סעיף ג), גם את הרשימה החדשה *LIST E2* ניתן להציג כאיחוד של שתי תת-רשימות: תת-הרשימה *E2* הבנויה מאותם זוגות ביטויים (w, w'), שבהם w הוא מן השמות והכינויים המיוחדים, ותת-הרשימה *E'2* הבנויה מאותם זוגות ביטויים (w, w'), שבהם w הוא הכינוי רבי פלוני.
4. יש לשים לב לכך, כי *E'2* אמור להיות דומה מאד ל- *L'2* באשר מדובר באותם אישים (למעט שינויים בודדים). לכן, ידוע מראש (תת-סעיף 2 לעיל) כי *E'2* רק יקלקל את התוצאה. תמוה הדבר, שמבב"ק כללו תת-רשימה זו בניסוי. הדבר תמוה שבעתיים לאור טענתם המרכזית של מבב"ק הנוגעת ל- *L2* בלבד (ראו במבוא, סעיף ג5).
5. מבב"ק [5] מדדו את תת-הרשימה *E2* בספר בראשית, ומצאו (במבחן הפרמוטציות) כי המובהקות של הסיגנאל S_2 היתה חזקה יותר מזו של S_1. גם תוצאה זו היתה צפויה מראש ברמה גבוהה של סבירות לאור הדמיון בין *E2* ל- *L2*, ולאור היתרון הגדול במובהקות S_2 לעומת S_1 לגבי *L2* במבחן הפרמוטציות [1].
6. לכן, המדידה המעניינת היחידה לגבי תת-הרשימה *E2* בספר בראשית היא של הסיגנאל S_2. מבב"ק קיבלו רמת מובהקות של 0.017 עבור S_2 במבחן הפרמוטציות[3]. כאן נשוב ונמדוד בספר בראשית אותו סיגנאל עבור אותה תת-רשימה, אך במבחן הטקסטים.

ב. הטקסט:

הקטע *D*, שהוא ספר בראשית (*G*), הוא "הקטע [1, 78064] מן התורה" (1 הוא המספר הסדורי של האות הראשונה בתורה, ו- 78064 הוא המספר הסדורי של האות האחרונה בספר בראשית). נסמן (לפי

[3] מבב"ק הצליחו להציג תוצאה זו כלא-מובהקת ע"י שימוש בבחירותיהם התמוהות בתת-סעיפים 4 ו-5, דבר שהצריך לדבריהם תיקון בונפרוני (דהיינו, הכפלת התוצאה פי 4).

המבוא סעיף א1): $D = G = ([1, 78064], TORAH)$. ניסוי מבב"ק על *E2* נעשה בקטע זה, כמו הניסוי המקורי של ור"ר על *L2*.

ג. מדידת המדגמים:

1. כל אחד מן המדגמים [א], [ב] ו[ג] נמדד ב-G באותו אופן בדיוק כפי שנמדד המדגם המקורי ב[1]. כך נקבע ערכו הגולמי של הסיגנאל S_2 עבור כל מדגם.
2. כדי לקבוע את מובהקות הסיגנאלים, נעשתה השוואה לאוסף של 50,000 טקסטים T_j, שכל אחד מהם נוצר על ידי עירבוב אקראי של G, דהיינו, עירבוב של מלים בתוך הפסוקים (כמבואר במבוא סעיף ב), ונקבע ערכו של n_2, מספר המתחרים שלהם ערך $S_2^j \geq S_2$.
3. לפי n_2, חושבה מובהקות הסיגנאל S_2: $p_2 = n_2 / 50{,}000$.

התוצאות מוצגות בטבלה 1.

טבלה 1

מספר המתחרים n_2 שלהם ערך $S_2^j \geq S_2$ בתחרות ובה 50,000 מתחרים.

המדגם	*E2(ABCD)*	*E2(ABC)*	*E2(AB)*
ערכי S_2	3.489E+2	3.628E+2	1.533E+3
n_2	218	199	39
p_2	4.36E−3	3.98E−3	7.8E−4

יש לשים לב לכך שהתוצאות אינן בלתי תלויות. אדרבא, קיימת תלות חזקה בין תוצאות *E2(AB)* לבין תוצאות *E2(ABC)* ו-*E2(ABCD)*. התלות הנובעת מן העובדה הפשוטה, כי המדגם שבו הרבה תצורות תאריך מכיל את המדגם שבו מעט תצורות תאריך.

ד. חישוב המובהקות הכוללת של S_2 עבור *E2* בספר בראשית:

המובהקות החזקה ביותר במבחן הטקסטים נתקבלה עבור *E2(AB)* והיא $p_2 = 7.8\text{E}{-4}$. אומנם, זו לא היתה המדידה היחידה, ולכן צריך לקחת בחשבון שהיו עוד שתי מדידות. הדבר יכול להיעשות על ידי תיקון בונפרוני לפיו מכפילים את המובהקות הנ"ל פי 3 ומקבלים ערך של 2.34E−3. אולם תיקון בונפרוני נחשב שמרני, ובמיוחד במקרה כשלפנינו בו ישנה תלות חזקה בין התוצאות (עוד בעניין זה – בנספח סעיף ד).

אם מדרגים את כלל הטקסטים בתחרות עבור *E2(AB)*, הסיגנאל S_2 בספר בראשית מדורג במקום ה-40. השאלה העומדת לפנינו היא: מהו הסיכוי של טקסט להיות מדורג באקראי במקום 40-1 לפחות באחת משלוש התחרויות שבטבלה 1? – התשובה היא: סך כל הטקסטים השונים המאכלסים את 40 המקומות הראשונים בכל אחת משלושת התחרויות המסוכמות בטבלה 1, מחולק במספר הטקסטים שהשתתפו בתחרות. אם התחרויות היו בלתי תלויות, היינו מקבלים 40×3 טקסטים שונים המאכלסים את 40 המקומות הראשונים (או קרוב לזה). לעומת זאת, בגלל התלות בין המדגמים, כאן יש רק 69 טקסטים שונים כאלה. לכן, המובהקות הכוללת של S_2 עבור *E2* בספר בראשית היא:

$$\rho_2 = 69/50{,}000 = 1.38\text{E}{-3}$$

ה. לסיכום:

1. הסיגנאל S_2 של כל אחד משלושת המדגמים [א], [ב] ו[ג] נמצא מובהק בספר בראשית.
2. מובהקות הסיגנאל החזקה ביותר היא עבור מדגם [א] וערכה $p_2 = 7.8\text{E}{-4}$.
3. המובהקות הכוללת של S_2 עבור *E2* בספר בראשית היא: $\rho_2 = 1.38\text{E}{-3}$.

ו. הערות:

1. תוצאות אלה מרשימות שבעתיים אם נביא בחשבון את התערבות מבב"ק בהכנת הנתונים (ראו בראש הפרק, ובהרחבה ב[8]), באמצעותה הצליחו להשיג במבחן הפרמוטציות עבור תת-הרשימה *E2* מובהקות שאינה חזקה מ-0.017 עבור S_2.
2. על מנת לגמד את ערך התוצאה של 0.017 טענו מבב"ק [5], כי תוצאה זו מוטה לטובת ור"ר (!), וכי הגורם "החשוב ביותר" להטיית התוצאה הוא השימוש שנעשה במבחן הפרמוטציות (במתכונת שבוצעה ב[1]). ובכן, מלבד שטענתם בנידון חסרת שחר – הנה המדידות שלפנינו, שנערכו במבחן

הטקסטים (לגביו טענתם לא רלוונטית), מוכיחות להיפך: מבחן הפרמוטציות דווקא החליש את המובהקות (עוד בנספח, סעיף ב).

מדידה 2: מדידת הסיגנאל S_2 עבור תת-הרשימה $E2$ בספר "מלחמה ושלום".

א. הטקסט:

נסמן את הספר "מלחמה ושלום" [7] ב- $W\&P$. אחד מן הטקסטים ששימשו לניסויי ביקורת ב[1] היה הקטע $WP = (W\&P, [1, 78064])$. מבב"ק [5] השתמשו בו לצורך הדגמת פרודיה על עבודת ור"ר.

ב. מדידת המדגמים:

1. כל אחד מן המדגמים [א], [ב] ו[ג] נמדד ב-WP באותו אופן בדיוק כפי שנמדד המדגם המקורי ב[1]. כך נקבע ערכו הגולמי של הסיגנאל S_2 עבור כל מדגם.
2. כדי לקבוע את מובהקות הסיגנאלים, נעשתה השוואה לאוסף של 1000 טקסטים T_j, שכל אחד מהם נוצר על ידי עירבוב אקראי של WP, ונקבע ערכו של n_2, מספר המתחרים שלהם ערך $S_2^j \geq S_2$.
3. לפי n_2, חושבה מובהקות הסיגנאל S_2: $p_2 = n_2 / 1000$.

התוצאות מוצגות בטבלה 2.

טבלה 2

מספר המתחרים n_2 שלהם ערך $S_2^j \geq S_2$ בתחרות ובה 1,000 מתחרים.

המדגם	*E2(ABCD)*	*E2(ABC)*	*E2(AB)*
ערכי S_2	3.953	5.814	5.128
n_2	218	125	151
p_2	0.218	0.125	0.151

התוצאות בטבלה 2 מצביעות על חוסר מובהקות.

ג. מסקנות:

עבור תת-הרשימה $E2$:

1. יש הבדל מהותי בין ערכו הגולמי של הסיגנאל S_2 בספר בראשית לבין ערכו ב"מלחמה ושלום".
2. יש הבדל מהותי בין מובהקות הסיגנאל S_2 בספר בראשית לבין מובהקותו ב"מלחמה ושלום".

מדידה 3: מדידת תת-הרשימה $E2$ בתחום המורחב.

א. מה נמדוד:

במאמרנו [3] הסקנו, כי תחום ההצפנה של רשימת הנתונים $L2$ קרוב יותר לתחום המורחב, שהוא הקטע [1, 156128] מן התורה, מאשר לתחום המקורי, הקטע [1, 78064] מספר התורה. לכן, יש עניין רב לערוך ניסויים נוספים בנושא רשימת נתונים זו בתחום המורחב. כמה ניסויים כאלה אכן נערכו ובהצלחה רבה [4]. כאן עלה הרעיון לבדוק את תת-הרשימה $E2$, שאף היא ממש בנושא רשימת נתונים הנ"ל, בתחום המורחב.

ב. הטקסט:

הטקסט הוא הקטע [1, 156128] מן התורה: $D' = (TORAH, [1, 156128])$.
קטע זה נקרא ב[3] וב[4] "התחום המורחב", והוא כפול באורכו מספר בראשית.

ג. מדידת המדגמים:

1. כל אחד מן המדגמים [א], [ב] ו[ג] נמדד בקטע D' באותו אופן בדיוק כפי שנמדד המדגם המקורי ב[1]. כך נקבע ערכם הגולמי של הסיגנאלים S_1 ו- S_2 עבור כל מדגם.
2. נחשב את מדד ההגברה Q_i לגבי הסיגנאלים (ראו הסבר במבוא סעיף א5), וכן את מדד ההגברה המירבית $Q \equiv \max\{Q_1, Q_2\}$, ביחס לסיגנאלים המקבילים בקטע D (ספר בראשית).

התוצאות מוצגות בטבלה 3.

טבלה 3

E2(AB)		*E2(ABC)*		*E2(ABCD)*		המדגם
1	2	1	2	1	2	i
3.222E+4	2.644E+5	2.922E+3	1.138E+4	1.111E+3	4.313E+3	ערכי S_i
835.8	172.5	39.4	29.6	15.0	12.4	Q_i
835.8		39.4		15.0		Q

ד. מובהקות המדגמים:

1. כדי לקבוע את מובהקות הסיגנאלים, נעשתה השוואה לאוסף של 50,000 טקסטים T_j שכל אחד מהם נוצר על ידי עירבוב אקראי של D', ונקבע ערכו של n_i, מספר המתחרים שלהם ערך $S_i^j \geq S_i$.
2. לפי n_i, חושבו מובהקויות הסיגנאלים S_i: $p_i = n_i / 50{,}000$.

התוצאות מוצגות בטבלה 4.

טבלה 4

מספר המתחרים n_i שלהם ערך $S_i^j \geq S_i$ בתחרות ובה 50,000 מתחרים.

E2(AB)		*E2(ABC)*		*E2(ABCD)*		המדגם
S_1	S_2	S_1	S_2	S_1	S_2	סיגנאל
4	0	71	18	173	39	n_i
8E−5	< 2E−5	1.42E−3	3.6E−4	3.46E−3	7.8E−4	p_i

ה. סיכומי תוצאות:

1. המובהקות הכוללת עבור *E2(ABC)* בתחום המורחב:

כאן חישבנו את המובהקות הכוללת באמצעות ספירת כלל הטקסטים השונים המאכלסים את 19 המקומות הראשונים בכל אחת משתי התחרויות: עבור הסיגנאלים S_1 ו- S_2. אם התחרויות היו בלתי תלויות, היינו מקבלים 19×2 טקסטים שונים כאלה (או קרוב לזה). לעומת זאת, בגלל התלות בין הסיגנאלים, כאן יש רק 31 טקסטים שונים כאלה. לכן, המובהקות הכוללת עבור *E2(ABC)* בתחום המורחב היא:

$$\rho = 31/50{,}000 = 6.2\text{E}{-4}$$

2. המובהקות הכוללת עבור *E2* בתחום המורחב:

המובהקות החזקה ביותר במבחן הטקסטים נתקבלה עבור הסיגנאל S_2 של *E2(AB)* והיא $p_2 < 2\text{E}{-5}$. גם כאן ישנה תלות בין התוצאות, ולכן רק 5 טקסטים שונים הגיעו לראש הדירוג בשש התחרויות. על כן, המובהקות הכוללת עבור *E2* בתחום המורחב היא:

$$\rho = 5/50{,}000 = 1\text{E}{-4}$$

הערה: להערכתנו, בהרצה של 1,000,000 טקסטים צפויה המובהקות הכוללת להתחזק בערך פי 10.

ו. לסיכום:

1. הסיגנאלים של כל אחד משלושת המדגמים [א], [ב] ו[ג] התחזקו בתחום המורחב.
2. הסיגנאלים של כל אחד משלושת המדגמים [א], [ב] ו[ג] נמצאו מובהקים בתחום המורחב.
3. המובהקות החזקה ביותר היא בסיגנאל S_2 עבור מדגם [א] וערכה $p_2 < 2\text{E}{-5}$.
4. המובהקות הכוללת של הסיגנאלים בתחום המורחב היא: $\rho = 1\text{E}{-4}$.

מדידה 4: מדידת תת-הרשימה *E2* בספר "מלחמה ושלום" בקטע שאורכו כאורך התחום המורחב.

א. הטקסט:

הוא הקטע (*WP1* = ([1, 156128], *W&P*) מספר "מלחמה ושלום" [7]. בקטע זה השתמשנו לביקורת במאמרנו [3].

ב. מדידת המדגמים:

1. כל אחד מן המדגמים [א], [ב] ו[ג] נמדד בקטע *WP1* באותו אופן בדיוק כפי שנמדד המדגם המקורי ב[1]. כך נקבע ערכם הגולמי של הסיגנאלים S_1 ו- S_2 עבור כל מדגם.
2. כדי לקבוע את מובהקות הסיגנאלים, נעשתה השוואה לאוסף של 1000 טקסטים T_j שכל אחד מהם נוצר על ידי עירבוב אקראי של *WP1*, ונקבע ערכו של n_i, מספר המתחרים שלהם ערך $S_i^j \geq S_i$.
3. לפי n_i, חושבו מובהקויות הסיגנאלים S_i: $p_i = n_i / 1000$.

התוצאות מוצגות בטבלה 5.

טבלה 5

מספר המתחרים n_i שלהם ערך $S_i^j \geq S_i$ בתחרות ובה 1000 מתחרים.

E2(AB)		*E2(ABC)*		*E2(ABCD)*		המדגם
S_1	S_2	S_1	S_2	S_1	S_2	סיגנאל
1.197	7.092	1.066	2.597	1.030	1.661	ערכי S_i
729	117	849	300	904	475	n_i
0.729	0.117	0.849	0.300	0.904	0.475	p_i

התוצאות בטבלה 5 מצביעות על חוסר מובהקות.

ג. לסיכום:

עבור תת-הרשימה *E2*:

1. יש הבדל מהותי בין ערכם הגולמי של הסיגנאלים בתחום המורחב לבין ערכם ב"מלחמה ושלום".
2. יש הבדל מהותי בין מובהקות הסיגנאלים בתחום המורחב לבין מובהקותם ב"מלחמה ושלום".

מסקנות

א. קיים סיגנאל לא-טריוויאלי של רשימת "החיקוי" – תת-הרשימה *E2* – בספר בראשית.
ב. הסיגנאל של רשימת "החיקוי" מתגבר במעבר לתחום המורחב.
ג. מובהקות הסיגנאל של רשימת "החיקוי" בתחום המורחב חזקה מאד.
ד. כפי שכתבו מבב"ק עצמם, לרשימת "החיקוי", שנוצרה על ידי מומחה מטעמם, יש חשיבות גדולה מאד: התוצאות א-ג מוכיחות כי יש כאן תופעה אמיתית של הצפנת נתונים, וכי התוצאה המקורית של ור"ר איננה תולדה של מניפולציה של שמות וכינויים.

נספח

א. המידות הכוללות לקירבה:

1. בניסוי המקורי של ור"ר ([1], עמ' 436) הוגדרו ארבע "מידות כוללות לקירבה" P_1, P_2, P_3 ו- P_4. שם ([1], עמ' 431) מבואר כי:
   - מידות P_1 ו-P_2 הוגדרו כשני סטטיסטים שונים לגבי הרשימה *LIST 2*.
   - כאשר מידות P_1 ו-P_2 מיושמות לגבי תת-הרשימה *L2*, הן מסומנות כ- P_3 ו- P_4.

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

היישום שלהם על רשימת הנתונים השלמה נותן את P_1 ו- P_2:

$$P_1 \equiv \boldsymbol{P_1}(LIST\ 2),\quad P_2 \equiv \boldsymbol{P_2}(LIST\ 2)$$

והיישום שלהם על תת-הרשימה *L2* נותן את P_3 ו- P_4:

$$P_3 \equiv \boldsymbol{P_1}(L2),\qquad P_4 \equiv \boldsymbol{P_2}(L2)$$

בניסוי הנוכחי לא נשתמש עוד באינדקסים 3 ו-4, וכאשר נשתמש באופראטורים $\boldsymbol{P_1}$ ו- $\boldsymbol{P_2}$ יובהר תמיד לגבי איזו רשימה הם מיושמים.

2. להקל על הקורא נביא כאן את הגדרת P_1 ו- P_2 מ[1].

(א) הגדרת מידת "הנטייה הכוללת לקרבה" 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 המכוילת" 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

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

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. אומנם, איננו עושים כאן כל שימוש בהנחה של אחידות ואי-תלות. לכן, למרות ש-P_1 מכויל כהסתברות, הוא משמש רק כמדד סידורי. 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 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.

אנו מחשבים את המכפלה $\prod c(w,w')$ ואחר כך אנו מגדירים

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

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

אומנם, איננו עושים כאן כל שימוש בהנחה של אחידות ואי-תלות. לכן, למרות ש-P_2 מכוילת כהסתברות, היא משמשת רק כמדד סידורי.

ב. מדידת מובהקות הסיגנאל

1. כאשר ור"ר [1] ביצעו בספר בראשית מדידות של רשימות הנתונים, הם מדדו את האפקט הכולל באמצעות מידות כוללות P_i. כדי למדוד את מובהקות התוצאות היה צריך לחזור על אותן מדידות בטקסטים רבים "דומים". דבר זה לא היה בהישג ידם של ור"ר לפני 28 שנים (מבחינת כמות החישובים)[4]. הוצע להם, לערוך מבחן פרמוטציות, אשר פרטיו סוכמו בין פרסי דיאקוניס וישראל אומן, ואושרו לפני עריכת הניסוי על ידי ארבעה סטטיסטיקאים נודעים נוספים. מבחן זה בוצע ב-[1]. מבחן הפרמוטציות מעריך את מובהקות הסטטיסטים P_i שנתקבלו בספר בראשית עבור רשימת הנתונים המקורית (מפגשים של שמות אישים עם תאריכי הלידה ו/או הפטירה שלהם), באמצעות השוואה לערכי הסטטיסטים P'_i המתקבלים בספר בראשית עבור הרבה רשימות נתונים "מעורבבים" (מפגשים של אותם שמות אישים עם תאריכי הלידה ו/או הפטירה של אישים אחרים מן הרשימה המקורית).

2. כחמש וחצי שנים לאחר ביצוע מבחן הפרמוטציות המתואר ב[1], סבר אליהו ריפס שהתקדמותם המהירה של אמצעי החישוב מאפשרת לשוב לרעיון הפשוט ביותר של מדידת המובהקות: באמצעות רנדומיזציה של טקסטים, דהיינו, באמצעות השוואה לטקסטים רבים "דומים". דרך הביצוע מתוארת במבוא למאמר, סעיף ב. מבחן מסוג זה יכונה להלן "מבחן טקסטים".

---

[4] להרצת רשימה בגודלה של *LIST2* נדרשו כשלושה שבועות.

3. השוואה בין מבחן הפרמוטציות למבחן הטקסטים – שיקול כללי:
כל אחד משני המבחנים מודד דבר שונה:

- מבחן הפרמוטציות מודד עד כמה "מוצלחים" המפגשים של שמות האישים עם תאריכי הלידה ו/או הפטירה שלהם יותר מן המפגשים של אותם שמות אישים עם תאריכים אחרים (שהם תאריכי לידה/פטירה של אישים אחרים מאותה רשימת נתונים).
- מבחן הטקסטים מודד מה הסיכוי לקבל במקרה P_i כה קטנים. דהיינו: מה הסיכוי שהצטברו במקרה כל כך הרבה מפגשים "מוצלחים".

במלים אחרות: מבחן הטקסטים מודד את מובהקות הכמות המצטברת של מפגשים "מוצלחים", ואילו מבחן הפרמוטציות מודד את המובהקות של "ספציפיות הצופן".
לכן אנו מצפים אפריורי להבדל בתוצאות שני המבחנים עבור אותה רשימת נתונים, שהרי הם מודדים דברים שונים.

4. דוגמא:
רשימה הנתונים *BM2* שנמדדה בעבודת מבב"ק [5] בספר "מלחמה ושלום", הוכנה על ידי מניפולציה מוצהרת. המניפולציה נעשתה כדי להשיג מובהקות חזקה במבחן הפרמוטציות. ואכן, הושגה מובהקות חזקה מאד עבור הסטטיסטי הנמדד:

$$p = 7E\text{-}7.$$

לעומת זאת, כאשר מודדים את המובהקות של אותו סטטיסטי במבחן טקסטים, מתברר כי ב-12 מתוך 10,000 טקסטים "דומים" (שנוצרו מערבוב אקראי של הטקסט המקורי) – הושגה תוצאה חזקה יותר. לכן, המובהקות חלשה בהרבה:

$$p = 1.2E\text{-}3.$$

אכן, קיים הבדל גדול מאד בין התוצאות.

5. המבחן המתאים למצבו העכשווי של מחקר הצפנים בתורה:
המחקר המדעי של הצפנים בתורה, עד כמה שבא לידי ביטוי במדידת רשימות גדולות של נתונים, עסק עד כה בהוכחת קיומו של צופן מסויים. הוא לא עסק באיפיון מדוייק או בפיענוח של הצפנים. הדבר הוצהר במפורש, בהקדמת ור"ר [1] (בעמוד הראשון של המאמר. בעניין זה ראו גם [2] עמ' 52-57). לכך מתאים דווקא מבחן הטקסטים, ולא מבחן פרמוטציות המודד את המובהקות של "ספציפיות הצופן". (ור"ר השתמשו במבחן פרמוטציות רק באין ברירה, כמבואר לעיל).

6. השוואה בין מבחן הפרמוטציות למבחן הטקסטים – שיקול ספציפי:
בדרך כלל משתמשים במבחן פרמוטציות במקרה הבא:
ישנה קבוצה **א** שלגביה השערת המחקר טוענת טענה מסוימת, וקבוצת ביקורת **ב** שלגביה אין השערת המחקר טוענת דבר. לפי השערת האפס אין תופעה גם ב**א** וגם ב**ב**. כלומר יש ויכוח בין השערת האפס להשערת המחקר ביחס לקבוצה **א** ויש הסכמה ביחס לקבוצה **ב**.

(א) בוחרים סטטיסטי לאפיון התופעה.

(ב) מגרילים באקראי פריטים (בכמות של פריטי **א**) מקבוצה **ג**, שהיא האיחוד של **א** ו**ב**, ומודדים את הסטטיסטי לגביהם.

(ג) חוזרים על (ב) פעמים רבות, ורואים את המיקום של קבוצה **א** בהתפלגות ערכי הסטטיסטי.

לעומת זאת, מבחן הפרמוטציות שהוצע לור"ר ב[1] היה שונה.

המדגם המקורי הוא קבוצה **א**. במבחן הפרמוטציות, אנו מצמידים לכל אדם תאריך של חברו לרשימה. קבוצת כל הזוגות של שמות אישים שלהם מוצמדים תאריכים של אחרים, אמורה להיות קבוצת הביקורת **ב**.

אבל כאן, השערת המחקר אינה מסכימה שלגבי **ב** ודאי אין תופעה. וכי אפשר לטעון (למשל) כי י"ד ניסן, יום הולדתו של הרמב"ם, אינו קשור להראב"ד? וכי הראב"ד לא עשה (או חידש) דברים מהותיים בערב פסח בכל ימי חייו? וכי יום כ"ו כסלו (נר שני של חנוכה), יום מותו של הראב"ד, לא היה משמעותי בחיי הרמב"ם? ומי בכלל יכול לטעון, כי תאריך מסוים לא קשור לאישיות מסוימת?!

לכן, ייתכן מאד שהמדגם המקורי מתחרה עם מדגמים משובשים המכילים זוגות מקבוצה **ב** האמורים להצליח אם קיימת התופעה, והדבר צפוי להחליש את המובהקות הנמדדת.

7. במבחן הטקסטים לא קיימת בעיה זו.

ג. הרשימה שפרסמו מבב"ק

הרי הרשימה שפורסמה על ידי מבב"ק בשמו של עמנואל (יחד עם הערותיהם). במאמר הנוכחי נקראת רשימה זו *LIST E2*.

```
# Corrected second list, as prepared by Simcha Emanuel with
# knowledge of the rules and appellations in WRR(1986)

11 1 רביאברהמאבדמנרבונא אברהמבניצחקמנרבונה אברהמבניצחקמנרבונא רביאברהמ
     בעלהאשכול האשכול הראבי הראביאבד ראבי ראביאבד רביאברהמאבדמנרבונה
 חשון/כ/

3 1 סיון/יג/  אברהמיצחקי יצחקי רביאברהמ

3 1 תשרי/יב/  רביאברהמהמלאכ אברהמבנדבבערממעזריטש רביאברהמ

5 0 בעלצרורהמור צרורהמור אברהמסבע סבע רביאברהמ

3 1 כסלו/כב/ אליעזראשכנזי אשכנזי רביאליעזר

3 1 תשרי/ז/  דודאופנהיים אופנהיים רבידוד

2 0 דודבנאברהמהנגיד רבידוד

5 1 טבת/כח/  הדנ דנ דודניטו ניטו רבידוד

5 1 ניסן/ו/  חיימאבואלעפיא חיימאבולעפיה אבואלעפיא אבולעפיה רביחיים

5 1 אלול/יט/  בעלכנסתהגדולה כנסתהגדולה חיימבנבנשת בנבנשת רביחיים

4 1 שבט/יב/  רביחיימבעלהנס חיימכפוסי כפוסי רביחיים

7 1 ניסן/יג/  המהרחש הרחש מהרחש רחש חיימשבתי שבתי רביחיים

5 1 טבת/א/  בעלחוותיאיר חוותיאיר יאירחיימבכרכ בכרכ רבייאירחיים [see note]

4 0 רבייהודההחסיד רבייהודהחסיד יהודהסגלמשידלוב רבייהודה

2 2 תשרי/יא/ טבת/ט/  יהוספבנשמואלהנגיד רבייהוסף

4 1 אב/כז/  בעלמגינישלמה מגינישלמה יהושעבניוספמקראקא רבייהושע

7 1 תמוז/יד/ המהרימט המהריט מהרימט מהריט יוספבנמשהמטרני יוספבנמשהמטראני רבייוסף

5 1 איר/י/  בעלפרימגדים פרימגדים יוספתאומים תאומים רבייוסף

3 1 איר/א/  יעקבבירב בירב רבייעקב

5 1 שבט/כו/  יעקבישראלחגיז יעקבישראלחאגיז חגיז חאגיז רבייעקבישראל

4 1 אלול/כב/  המהריל מהריל יעקבהלוימולין רבייעקב

5 3 סיון/יה/ סיון/טו/ ניסן/ל/  היעבצ יעבצ יעקבעמדין עמדין רבייעקב

7 1 רביאיציקלמהאמבורג רבייצחקמהאמבורג יצחקהורוביצ הורוביצ רבייצחק
    איר/ו/  רביאיציקלמהמבורג רבייצחקמהמבורג

5 1 שבט/ב/  מנחממנדלקרוכמל מנחממנדלקראכמאל קרוכמל קראכמאל רבימנחממנדל

5 2 תשרי/יו/ תשרי/טז/  הרמז רמז משהזכות זכות רבימשה
```

```
5 1 טבת/יב/  בעלפנימשה פנימשה משהמרגלית מרגלית רבימשה

5 1 אדרא/א/  בעלגדוליתרומה גדוליתרומה עזריהפיגו פיגו רביעזריה

5 3 בעלמשנתחסידים משנתחסידים רפאלעמנואלחיריקי ריקי רבירפאלעמנואלחי
 תמוז/יה/ תמוז/טו/ אדר/א/

5 1 שבט/י/  הרשש רשש שלומשרעבי שרעבי רבישלום

4 1 תמוז/כא/  בעלמרכבתהמשנה מרכבתהמשנה שלמהבנמשהמחלם רבישלמה

4 1 תמוז/כא/  רביאליהבעלשם אליהלואנצ לואנצ רביאליה

2 0 חנוכבנמשה רביחנוכ

10 1 רבימאיראש מאיראייזנשטאט מאיראייזנשטט אייזנשטאט אייזנשטט רבימאיר
      סיון/כז/  בעלפנימאירות פניממאירות המהרמאש מהרמאש

Note:  After publication of our paper, Dr Emanuel realised that the
appellation יאיר חוות is properly spelt יאיר חות.  Correcting the
error makes hardly any difference to the results of the experiment.
The overall significance levels given in our paper as 0.066 and 0.108
are still true.
```

ד. על תיקון בונפרוני.

בעבודת ור"ר [1] נעשה שימוש בפרוצדורה של בונפרוני. לפי בונפרוני אם α הוא הסף המבוקש ויש m סטטיסטים בניסוי, הרי צריכה להיות תוצאה של (לפחות) אחד הסטטיסטים $p \le \alpha / m$.

השימוש בפרוצדורה של בונפרוני נחשב כשמרני עד שמרני מאד (חיפוש במרשתת מגלה כי צמד המלים בונפרוני ו"שמרני" שכיח. יש המשתמשים אפילו בתואר "דרקוני"). במיוחד התיקון נחשב שמרני כאשר קיימת תלות בין הסטטיסטים.

1. בעבודת ור"ר היו 4 סטטיסטים. תוצאותיהם המדורגות במבחן הפרמוטציות:

4/1,000,000 5/1,000,000 453/1,000,000 570/1,000,000

לפי תיקון בונפרוני: $p = 4\times 4/1{,}000{,}000 = 16/1{,}000{,}000$.

למעשה, הסטטיסטים הללו תלויים זה בזה תלות חיובית חזקה. נטפל בכך תוך ניצול התוצאות של מבחן פרמוטציות שבידינו.

2. נדגים זאת:

הסטטיסטים P_2 ו- P_4 תלויים חזק זה בזה (P_4 הוא בדיוק P_2 המיושם לגבי תת-רשימה של נתונים כפי שמבואר לעיל בסעיף א).

(א) הדירוג של P_2 היה חמישי מתוך מיליון פרמוטציות. ארבע פרמוטציות הקדימו את המדגם המקורי:
#300087, #620851, #788884, #808836.

(ב) הדירוג של P_4 היה רביעי מתוך מיליון פרמוטציות. שלוש פרמוטציות הקדימו את המדגם המקורי:
#777442, #788884, #808836.

אם נסכם שני סטטיסטים אלה לפי בונפרוני נקבל: $p = 2\times 4/1{,}000{,}000 = 8/1{,}000{,}000$.

אבל אם נשאל: כמה פרמוטציות שונות הגיעו לדירוג 4-1 מתוך 1,000,000 פרמוטציות, לפחות באחד משני "המירוצים"? – התשובה לכך: 6 (כולל המדגם המקורי), והמובהקות היא לא יותר מ-6/1,000,000!

3. אם נסכם באופן דומה לגבי כל ארבעת הסטטיסטים, נקבל כי המובהקות היא 12 למיליון ולא 16 למיליון (כפי שחושב ב[1]).

4. בגוף המאמר השתמשנו בדרך סיכום זו לקבל מובהקויות כוללות.

הכרת תודה

החישובים נעשו באמצעות תוכנה של יעקב רוזנברג; ישֵׁרון יצחק לוי ומיכל לוי עזרו בהכנת הקטע שנוסף ל"מלחמה ושלום"; שלום סרברניק העלה, כדרכו, רעיונות מועילים; - תודה לכולם.

ביבליוגרפיה

1. עבודת ור"ר:

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

נמצא גם באתר "צופן בראשית" בקישור:

http://www.torahcode.co.il/pdf_files/pub/wrr.pdf

2. ד. ויצטום. "על הרמז בדילוג השווה: מדידה חדשה של מדגם גדולי חכמי התורה." בד"ד, חוברת 7, קיץ התשנ"ח. נמצא גם באתר "צופן בראשית" בקישור:

http://www.torahcode.co.il/pdf_files/pub/bdd.pdf

3. ד. ויצטום, סיגנאל של מערכת צפני *ELS*: דעיכה מול הגברה, התשע"ה. נמצא באתר הנ"ל בקישור:

http://www.torahcode.co.il/pdf_files/research/hagbara1.pdf

4. ד. ויצטום, סיגנאל של מערכת צפני *ELS*: ניסויים חדשים ב"תחום המורחב", התשע"ה. נמצא באתר הנ"ל בקישור:

http://www.torahcode.co.il/pdf_files/research/murchav1.pdf

5. עבודת מבב"ק:

McKay, B. D., Bar-Natan, D., Bar-Hillel, M. and Kalai, G. (1999). Solving the Bible Code puzzle. Statist. Sci. 14 No. 2 150-173.

6. בדואר אלקטרוני מיום 15 במאי שעה 16:40:13, שנת 1997 (למניינם).

7. "מלחמה ושלום" מאת ל. נ. טולסטוי. תרגום לעברית: לאה גולדברג (1953), ספריית פועלים, מרחביה.

8. ד. ויצטום, הוכחה סטטיסטית חדשה לקיום צופן בספר בראשית, התשס"א. נמצא באתר הנ"ל בקישור:

http://www.torahcode.co.il/eman_heb.htm

ראו עוד:

http://www.torahcode.co.il/pdf_files/oppose/ema1.pdf

9. ד. ויצטום, מדידת מובהקות במחקר הצפנים בתורה: תוצאות מדידות במבחן הטקסטים עבור עבודות המודדות מפגשי מד"שים עם רצפי אותיות בטקסט, התשע"ז. נמצא באתר הנ"ל בקישור:

http://www.torahcode.co.il/pdf_files/research/texts2.pdf

10. הספר:

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