ב״ה

מבחן הראנדומיזציה

מאת דורון ויצטום

הניסוי אשר נדרשו WRR לבצע, הוא מן הסוג הנקרא בעגה המקצועית ״מבחן ראנדומיזציה״. מבחן הראנדומיזציה הוא מבחן מובהקות מעניין ופשוט יחסית. מבחן המובהקות נועד לבדוק האם אכן מידות ״הנטייה הכוללת לקִרבה״ (הוגדרו ב״קירבה כוללת״) מקבלות ערכים ״נמוכים באופן חריג״ לגבי המדגם הנבדק. המבחן ניתן ליישום באופן דומה לגבי המדגם השני, לגבי המדגם הראשון ולגבי מדגמים אחרים בעלי אותו מבנה בסיסי. הגדרת המקרה הכללי למבחן זה מובאת להלן בנספח. כאן אתאר את יישום המבחן לבדיקת המדגם השני (ראה ״המדגם השני״), עבורו תוכנן.

בצורה ציורית אפשר לתאר כך את מבחן המובהקות למִדגם השני:

לכל אישיות מן המדגם נכין שתי מעטפות נפרדות: האחת – ״מעטפת השמות״ – בה נשים את כל השמות והכינויים של אישיות זו, והשניה – ״מעטפת התאריכים״ – בה נשים את תאריכי הלידה והפטירה שלה. סך הכל 32 מעטפות שמות ו-32 מעטפות תאריכים – כמספר האישים במדגם. ישנם אופנים רבים מאד לפיהם אפשר לצמד מעטפת תאריכים לכל מעטפת שמות. הצימוד הפשוט ביותר הוא כמובן זה, שבו מצמדים לכל מעטפת שמות את מעטפת התאריכים של אותה אישיות. אבל אפשר, למשל, גם לצמד לכל מעטפת שמות של אישיות מסוימת את מעטפת התאריכים של האישיות הבאה אחריה ברשימה. ואפשר גם לחשוב על צימודים מסובכים הרבה יותר.

כל צימוד קובע ״מִדגם״ באופן הבא: הצימוד מתאים מעטפת תאריכים לכל מעטפת שמות. כל צמד מעטפות כזה קובע קבוצה של זוגות ביטויים, אשר הביטוי הראשון בכל זוג הוא שם (או כינוי) ממעטפת השמות, ואילו הביטוי השני בו הוא תאריך הלקוח ממעטפת התאריכים. **סך כל זוגות הביטויים הללו, מכל צמדי המעטפות בצימוד המסוים, מהווה ״מדגם״ חדש.**

סך כל הצימודים האפשריים הוא !32 (נשתמש – כמקובל – בסימון ״!32״ לציון המכפלה: $32\times31\times...3\times2\times1$. מספר עצום – קצת יותר מ-260 מיליון מיליארד מיליארדים). על ידי צימודים אלה נוצרים !32 מדגמים, אשר **אחד** מהם הוא המדגם המקורי

(הנוצר מן הצימוד הפשוט, שבו מצמדים לכל מעטפת שמות את מעטפת התאריכים של אותה אישיות), והיתר הם **מדגמים משובשים** (משום שהם כוללים זוגות שבהם מצומדים שמות אישים לתאריכים לא להם). אפשר למדוד את מידת "הנטייה הכוללת לקִרבה" לגבי כל מדגם כזה. כך נקבל !32 מספרים, אותם נסדר לפי הסדר הרגיל של המספרים הממשיים.

אם התכונה שאנו מודדים היא אקראית, אזי למספר – שהוא ערכה של מידת "הנטייה הכוללת לקִרבה" במדגם המקורי – סיכוי שווה לתפוס כל אחד מ- !32 המקומות בסידור המספרים הזה. **זו השערת האפס שלנו**. נשים לב, שהשערת האפס, ומבחן המובהקות הנגזר ממנה, אינם עושים כל שימוש בשיקולים שהנחו אותנו בהגדרת "מידת הקרבה המכוילת" ומידות "הנטייה הכוללת לקרבה", שלפיהם יש להן משמעות הסתברותית (כפי שהוסבר ב"מדידת מפגשים" וב"קירבה כוללת"). לכן, מבחן המובהקות (אשר אתאר מיד) מהווה מבחן של **"קופסה שחורה"**.

המספר !32 הוא, כאמור, עצום ומשום כך אי אפשר לחשב את כל !32 המספרים הנזכרים לעיל. לכן, עלינו להסתפק בדגימה. כדי לחשב את רמת המובהקות הסטטיסטית, נערוך את המבחן הבא: נשים את מעטפות השמות בקלפי אחת, ואת מעטפות התאריכים – בקלפי אחרת. נערבב היטב את המעטפות שבקלפיות. עתה נוציא מעטפת שמות אחת מן הקלפי הראשונה, ולעומתה מעטפת תאריכים אחת מן הקלפי השניה. צמד המעטפות שעלה קובע קבוצה של זוגות ביטויים (כפי שהוסבר לעיל). עתה, נוציא מעטפה נוספת של שמות מן הקלפי הראשונה, ומעטפת תאריכים – מן השניה. נקבל עוד קבוצת זוגות ביטויים. כך נמשיך להגריל מעטפות, עד שיִכלו המעטפות מן הקלפיות. אחרי שהוצאנו את צמד המעטפות ה-32, האחרון, מצרפים את כל זוגות הביטויים שנקבעו על ידי 32 צמדי המעטפות, אשר יהוו "מדגם משובש" שנתקבל באמצעות **צימוד אקראי.**

נחזיר את המעטפות לקלפיות, נערבבן היטב ונחזור על התהליך 999,999 פעמים – כך נקבל 999,999 מדגמים משובשים. כפי שתיארתי לעיל, לכל מדגם כזה מחשבים את ערכה של מידת "הנטייה הכוללת לקרבה". יחד עם המידה של המדגם המקורי יש לנו בסך הכל 1,000,000 מספרים.

במבחן המובהקות עלינו לערוך עתה "תחרות" בין 1,000,000 המדגמים: באיזה מהם המפגשים "קרובים יותר" – כלומר, למי מהם שייכת מידת "הנטייה הכוללת לקִרבה" בעלת הערך הנמוך ביותר. [כפי שהוסבר ב"קירבה כוללת", אם ערכה של מידת "הנטייה הכוללת לקִרבה" עבור *מדגם א׳* **נמוך** מערכה עבור *מדגם ב׳*, פירושו של דבר, שהמפגשים *במדגם א׳* **"קרובים יותר"** מאלה *שבמדגם ב׳*.]

לשם כך, אנו מדרגים את 1,000,000 המספרים לפי סדר המספרים הממשיים: *בראש הדירוג – הערך המספרי הנמוך ביותר*. עתה נגדיר את "הדירוג" של המדגם המקורי מתוך 1,000,000 המדגמים המתחרים: זהו מספר המדגמים שערך מידת "הנטייה הכוללת לקרבה" שלהם *אינו עולה* על זה של המדגם המקורי[1].

לחישוב המובהקות נותר עתה רק צעד קטנטן. מ"השערת האפס" המנוסחת לעיל נובע, ש"הדירוג" של המדגם המקורי, המחולק במספר המתחרים (במקרה שלנו – 1,000,000) היא היא ההסתברות המבוקשת: ההסתברות, שערכה של מידת "הנטייה הכוללת לקִרבה" של המדגם המקורי *כה נמוך*.

נספח

בנספח זה נטפל בכמה נושאים הנוגעים לראנדומיזציה. בסעיף א' נגדיר את המקרה הכללי של מבחן הראנדומיזציה עבור מדגמים מן הצורה של המדגמים הגדולים – הראשון והשני. בסעיף ב' נביא את הפרטים הנוגעים לראנדומיזציה שבוצעה עבור המדגם השני. בסעיף ג' יוצגו הפרטים הנחוצים ליצירת הטקסטים R, U, V ו- W ששימשו כביקורת בניסויים.

א. מבחן ראנדומיזציה למִדגמים מטיפוס P:

המִדגמים של גדולי חכמי התורה, ששִמשו לעריכת שני הניסויים הגדולים, ניתנים להצגה סכמתית בצורה הבאה:

$$\begin{array}{ccc} A_1 & \longleftrightarrow & B_1 \\ A_2 & \longleftrightarrow & B_2 \\ \vdots & & \vdots \\ A_n & \longleftrightarrow & B_n \end{array}$$

כאשר A_i ו- B_i הן קבוצות של ביטויים, כך שהביטויים ב- A_i קשורים מושגית לאלה שבקבוצה B_i. המִדגם כולל את כל זוגות הביטויים (w,w') כך ש- w שייך ל- A_i ו- w' שייך ל- B_i, לכל $1 \le i \le n$.

[1] אם ישנם מדגמים, שערך מידותיהם שווה לערך מידת המדגם המקורי (כלומר, מקרה של "תיקו"), ייחשבו חציים כמקדימים את המדגם המקורי בדירוג.

למדגם כזה אנו קוראים מדגם מטיפוס P.

את n הקבוצות A_i אפשר להתאים ל- n הקבוצות B_i ב- $n!$ אופנים שונים (הנקראים פרמוטציות או צימודים). לכל פרמוטציה π מ- $n!$ הפרמוטציות, אפשר להתאים מדגם, הכולל זוגות ביטויים (w,w') כך ש-w שייך ל-A_i ו-w' שייך ל- $B_{\pi(i)}$, לכל $1 \le i \le n$.

עבור מדגם זה, מקבלת הסטטיסטיקה j (שהיא מידת "הנטייה הכוללת לקרבה") את הערך P_j^π.

את $n!$ ערכי P_j^π המתקבלים בפרמוטציות אלו, נסדר לפי הסדר הרגיל של המספרים הממשיים.

אם התכונה שבמעקב אינה אלא אקראית, הסיכוי ש- P_j (ערך הסטטיסטיקה j עבור המדגם המקורי) יאכלס כל אחד מ- $n!$ המקומות בסדור זה, הינו שווה. זו היא השערת האפס שלנו.

כדי לחשב את רמת המובהקות הסטטיסטית, אנו מגרילים M פרמוטציות אקראיות π של n עצמים, בדרך שנתאר בסעיף ב'. כל אחת מן הפרמוטציות π קובעת את הסטטיסטיקה P_j^π ; יחד עם P_j יש לנו $M+1$ מספרים. נגדיר את הדירוג של P_j בתוך $M+1$ המספרים הללו, כמספר ה- P_j^π שאינם קטנים מ-P_j ; אם P_j שווה ל- P_j^π אחרים, חציים של אלה ייחשב כ"מקדים" את P_j בדירוג.

נסמן ב- r_j את הדירוג של P_j, מחולק ב- $M+1$. בהשערת האפס, r_j הוא ההסתברות ש-P_j קרוב כל כך לראש הדירוג.

עד כאן התיאור הכללי של מבחן ראנדומיזציה עבור מדגם מטיפוס P. בקובץ "מדגם כותרת" יינתן התיאור הכללי של מבחן ראנדומיזציה עבור מדגם מטיפוס S (מדגם "כותרת").

ב. הראנדומיזציה בניסוי על המדגם השני של גדולי חכמי התורה:

כפי שכתבתי לעיל, מבחן הראנדומיזציה הראשון בוצע לגבי המדגם השני. בחירת 999,999 הפרמוטציות האקראיות של 32 האישים הכלולים בו, נעשתה באמצעות "אלגוריתם *P*" *של* קנות[2] [2] עמ' 125. המחולל הפסודו-רנדומלי ששימש כקלט לאלגוריתם זה, היה של *Turbo-Pascal 5.0* המיוצר על ידי *Borland Inter Inc.* "הזרע הסטטיסטי" הנדרש להפעלת המחולל הוא מספר שלם

[2] Knuth, D. E. (1969). The Art of Computer Programming, Vol. 2, Addison-Wesley, Reading, MA.

בן 32 ספרות בינאריות (כלומר, מספר שלם בעל 32 ספרות כאשר הוא נכתב לפי בסיס 2). הוא נקבע על ידי פרופסור אומן בעזרת השופטים (ראה בספר "צופן בראשית") להיות

01001 10000 10011 11100 00101 00111 11

(בהצגתו הרגילה, העשרונית, המספר הוא 1,277,674,143).

לפני עריכת הניסוי, דרש אחד השופטים לבצע אותו באמצעות 10 "זרעים" שונים, שכל אחד מהם ייצור סדרה של 100,000 פרמוטציות אקראיות. בכך הביע חשש מפני אפשרות של תלויות פנימיות במחולל הפסודו-רנדומלי. אולם הוחלט, שחשש זה רחוק והאפשרות הנ"ל היא בעלת סבירות נמוכה מאד. בכל זאת, לפני ביצוע הניסוי נדרשנו להסכים כי במקרה שנצליח בניסוי, והדרישה לערוך את הניסוי באמצעות 10 "זרעים" תועלה מחדש – נהיה חייבים לעשות זאת.
ואכן, לאחר שהצלחנו בניסוי, שבוצע כפי שנקבע מראש עם סדרה של 999,999 פרמוטציות אקראיות, נדרשנו (על ידי שופט של ה- *Statistical Science*, כתב העת אליו שלחנו את המאמר לפרסום) לחזור על הניסוי ולבצעו באמצעות 10 "זרעים" כנ"ל. תוצאות הניסוי בגירסתו זו היו אף טובות יותר. התוכנה שביצעה את הראנדומיזציה המתוארת בסעיף זה הוכנה בידי יעקב רוזנברג.

ג. יצירת טקסטים באמצעות ראנדומיזציה:

יצרנו ארבעה טקסטים על ידי ערבוב אקראי.

1. טקסט R, ששימש לביקורת, נוצר מפרמוטציה אקראית יחידה של 78,064 אותיות ספר בראשית. הפרמוטציה נוצרה בתהליך שתואר בסעיף הקודם, עם "זרע סטטיסטי" שנבחר להיות המספר העשרוני 10 (כלומר, השלם הבינארי 1010). טקסט R עורבב, אם כן, ברמה של האותיות.

2. טקסט W, ששימש לביקורת, נוצר על ידי פרמוטציה של המלים בספר בראשית, שנעשתה באותה צורה ובאותו "זרע". טקסט W עורבב, אם כן, ברמה של המלים: האותיות בכל מלה נשארו כסדרן.

3. טקסט V, ששימש לביקורת, נוצר מפרמוטציה אקראית יחידה של פסוקי ספר בראשית, שנעשתה באותה צורה ובאותו "זרע". טקסט V עורבב, אם כן, ברמה של הפסוקים: האותיות בכל פסוק נשארו כסדרן.

4. טקסט U, ששימש לביקורת, נוצר על ידי פרמוטציה של המלים בכל פסוק מספר בראשית, באותה צורה ובאותו "זרע", בעוד שהאותיות בכל מלה וכן הפסוקים - נשארו כסדרם. ביתר פירוט:

"אלגוריתם *P*" *של* קנות זקוק ל- $n-1$ מספרים אקראיים כדי לייצר פרמוטציה אקראית של n פריטים. המחולל הפסודו-רנדומלי שהשתמשנו בו, מייצר לכל "זרע סטטיסטי", שרשרת ארוכה של מספרים אקראיים. על ידי "הזרע" 1010 יצרנו שרשרת כזאת. ששת המספרים הראשונים בשרשרת נוצלו כדי לייצר פרמוטציה אקראית של שבע המלים המרכיבות את הפסוק הראשון בספר בראשית. שלושה-עשר המספרים הבאים (במקומות 7-19 בשרשרת הנ"ל) נוצלו כדי לייצר פרמוטציה אקראית של ארבע עשרה המלים המרכיבות את הפסוק השני, וכן הלאה.

הטקסטים הנ"ל נוצרו באמצעות תוכניות מחשב שהכין יעקב רוזנברג.